AD_____

Award Number: DAMD17-97-1-7130

TITLE: Computer-Assisted Visual Search/Decision Aids as a Training Tool for Mammography

PRINCIPAL INVESTIGATOR: Calvin F. Nodine, Ph.D.

CONTRACTING ORGANIZATION: University of Pennsylvania
Philadelphia, Pennsylvania 19104 -3246

REPORT DATE: July 2001

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 2001	Final (1 Jul 97 - 30 Jun 01)

4. TITLE AND SUBTITLE
Computer-Assisted Visual Search/Decision Aids as a Training Tool for Mammography

5. FUNDING NUMBERS
DAMD17-97-1-7130

6. AUTHOR(S)
Calvin F. Nodine, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pennsylvania
Philadelphia, Pennsylvania 19104-3246

E-Mail: nodine@oasis.rad.upenn.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Report contains color

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

14. SUBJECT TERMS

15. NUMBER OF PAGES
44

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

PROGRESS REPORT, 2000-2001, Year 4, DAMD17-97-1-7130 COMPUTER-ASSISTED VISUAL SEARCH/DECISION AIDS AS A TRAINING TOOL FOR MAMMOGRAPHY.
C.F. NODINE, PI 7/30/2001

(5) INTRODUCTION:

This project focuses on the perceptual training of diagnostic interpretation skills in mammography which are acquired mainly as a result of experience reading mammograms. The primary aim of this project is to develop a computer-assisted mammography training tool that will act as a surrogate mentor in aiding radiologists in making plausible diagnostic decisions. We propose to provide a computer aid that will interact with the radiologist immediately after image interpretation by providing systematic feedback about how the mammogram was searched for abnormalities and what features received prolonged visual attention indicating potential lesions during scanning. The eye-position parameter, visual dwell, is used to predict the locations of suspicious lesions on the mammogram (Krupinski, Nodine, Kundel, 1998; Nodine, Kundel, Mello-Thoms et al., 1999). The radiologist is then asked to re-examine the entire image including the highlighted areas, determine if any abnormal features are present, and re-evaluate the original diagnostic decision. This re-evaluation of suspicious regions with visual feedback provides a perceptually-guided basis for a plausible problem-solving diagnostic solution. We showed in 1990 (Kundel, Nodine, Krupinski, 1990) that computer-assisted visual search (CAVS) is effective in improving the detection of lung nodules, and Krupinski (1996) showed that visual dwell predicts the location of true and false, positive and negative decision outcomes. Our goal is to determine if CAVS improves the detection and interpretation of breast cancers.

(6) BODY:

(6.1) OBJECTIVES. The primary objective of the work this year was to test computer-assisted visual feedback system (CAVS) as a decision aid on radiology residents undergoing mammography training. The CAVS had become operational late last year. Early in 2001 we began recruiting radiology residents to carry out a formal mammography training experiment designed to evaluate the effectiveness of CAVS as a training tool for the study. This was the aim of Technical Objective 3. Task 6 under this objective was to test radiology residents in an experiment in which performance was compared with and without CAVS decision aid training. The feedback experiment was completed this Spring and analysis of the data has been completed fulfilling the aim of Task 7. We will now report on work completed from July 1, 2000 to July 1, 2001 based on the approved Statement of Work.

(6.2) TECHNICAL OBJECTIVE 3, MAMMOGRAPHY TRAINING EXPERIMENT.

(6.3) Task 6. Test radiology residents and CME fellows by assigning them to computer-assisted visual search (CAVS) or non-CAVS conditions and post test at

the end of training. This task was accomplished by using a crossover repeated-measures experimental design in which the same readers read the same test set of mammogram cases under both CAVS and Non-CAVS conditions. To control for practice effects, the order of the reading of the cases, and the assignment of CAVS or Non-CAVS was randomized across cases between the two reading sessions. To control for memory effects carrying over from one session to the other, one month separated the two sesions for each reader. We have programmed the ASL Model 4000 to monitor the observer's eye position relative to head motion for digital mammography displays during CAVS training. **TASK 6 Test radiology residents using CAVS as a training tool, COMPLETE.**

(6.4) Task 7. Analyze pre-post test differences using ROC analysis to measure performance. The effectiveness of CAVS as a training tool was evaluated by measuring decision performance with and without CAVS. A repeated measures analysis of variance was performed using area under the ROC curve (Az) as the dependent variable and CAVS vs. Non-CAVS conditions as the independent variable. The results will be reported below. **TASK 7 Analyze effects of CAVS vs. Non-CAVS training on decision accuracy, COMPLETE.**

(6.5) Role of Computer-Assisted Visual Search as a Training Tool for Mammography. Ten years ago we published a study showing that when radiology residents re-examined chest x-ray images with visual feedback, nodule detection performance was significantly higher than when they re-examined the same chest images without visual feedback (Kundel, Nodine, Krupinski, 1990). In this study, we found a significant 23% improvement in AFROC performance with visual feedback.

The rationale for feeding back regions of prolonged attention, indicated by eye-fixation clusters, is that, we have shown that false negative dwell times are more than twice as long in cumulative gaze duration as true negative dwell times (Kundel, Nodine, Krupinski. 1990). Cumulative gaze duration is defined by the sum of fixation durations within a cluster of fixations. Thus, feeding back regions on a chest or breast image where attention is concentrated, as indicated by cumulative gaze duration (dwell for short), but no decision is reported, may help identify fixated but unrecognized lesions. We have used 1000 ms as the dwell threshold for identifying feedback (FB) regions because it maximizes the likelihood of feeding back lesion-containing areas without feeding back numerous lesion-free areas (Krupinski, Nodine, Kundel, 1998). Because visual feedback may also prompt false-positive decisions, it is necessary to evaluate the effectiveness of visual feedback using a performance measure that adjusts for response bias. Both AFROC and LROC methods are ideal for this purpose and in addition, LROC gives us a measure of lesion-localization accuracy by differentiating between correctly localized TPs and incorrectly localized lesions on lesion-containing cases, called wrong lesions (WLs) (Swensson, 2000).

Materials and Methods

A crossover, repeated measures design was used in which 6 readers were assigned to experimental conditions by counterbalancing testing order and by randomizing the sequencing of 40 feedback and 40 non-feedback trials over two sessions separated by one

month in order to allow for forgetting. The 40 mammogram test set was presented on a high resolution 21" digital workstation (Clinton Electonics DS 5000L, Rockford, Il, 2560 x 2048). Each case was displayed in two views, craniocaudal (CC) in the left half of the screen and mediolateral oblique (MLO) in the right half of the screen.

Test Set

The test set of 40 mammogram cases, half containing malignant lesions and half lesion free, was presented to 6 readers, three radiology fellows and 3 radiology residents, having limited case reading experience (range 302-976 case readings prior to testing). A case consisted of two views of a single breast. All malignant-lesion cases but one had a single lesion in each view. One case had four lesions in each view. There were 13 cases with masses, 6 cases with microcalcifications, 1 case with architectural distortion and 20 lesion-free normal cases. The cases were selected by an experienced mammographer and represented "subtle lesion" mammogram cases. Each case was digitized using a 50 micron spot size by a Lumiscan 100 digitizer (Lumysis, Sunnyvale, CA).

Procedure

Each trial consisted of an initial overall-impression phase and final-decision phase. During the overall impression, readers were asked to evaluate CC and MLO views of each digitized breast case and decide whether or not it contained a malignant lesion. Eye-position was recorded using an ASL 4000 SU eye-head tracker (Applied Science Labs, Bedford, MA). When readers indicated they had finished evaluating each case, eye-position recording was terminated, and a menu was displayed. Readers used a mouse cursor to mark their overall-impression as either normal or abnormal and their initial decision confidence: high; medium; or, low. The final-decision phase followed during which eye-position data were analyzed and feedback regions calculated (regardless of whether feedback or control trial). During final decision phase, the reader was asked to re-evaluate the entire image including the highlighted feedback regions. If a malignant lesion was detected either from the overall impression, or newly discovered during the final decision, the reader clicked on the lesion location. This called up a menu and the reader indicated lesion type: mass; microcalifications; or, architectural distortion, and gave final decision confidence of malignancy of the lesion: high; medium; or, low.

Instructions

Readers were told that they could change their mind from initial overall-impression phase to final-decision phase and it was stressed that they should respond "abnormal" only if they consider that the case probably contained an malignant lesion. During the final-decision phase the reader was asked to localize malignant lesions in CC and MLO views (if possible) on cases called "abnormal" by the reader. However, it was emphasized that this did not preclude localization of newly discovered lesions during the final-decision phase on cases called "normal" initially. Conversely, readers could decide during final-decision phase that an initial "abnormal" decision was in error, and decide the case was free of a malignant lesion (i.e. "normal").

Data Analysis

The decision data for the initial overall-impression phase and final-decision phase were analyzed separately. In the overall impression, reader confidences in normal and abnormal overall decisions were used to construct a 2 x 6 ROC truth table. In the final decision, LROC analysis was applied to determine how many localized lesions matched true lesion locations as determined by an experienced mammographer. In scoring the final decision, if a localized lesion fell within 2.5 deg (1.65 cm) of a true lesion location, and was given the highest confidence for the case, it was scored as a correct response or true positive (TP) case. If a localized lesion was outside the 2.5 deg zone of a true lesion, and was given the highest confidence for the case, it was scored as an incorrect response or wrong lesion (WL). In tie cases where confidence was equal between a WL and a TP, the TP won for that case. Lesion-free cases were scored as correct or true negative (TN) if no lesions were localized, but if lesions were localized, the lesion with the highest confidence rating for the case was assigned as an incorrect response or false positive (FP). If a lesion was not localized in a lesion-containing case it was scored as a miss or false negative (FN) for the case.

The scoring of decision outcomes by case with associated confidence ratings were used to construct a 3 x 6 LROC truth table. Hits for decision outcomes in the truth table were defined as follows: To be counted a "hit" the cursor x,y coordinates had to be within 2.5 deg, of the lesion center. Minimum dwell for CAVS FB region was $\geq$ 1000 ms meaning that the cumulative fixation time of a group of fixations clustering on a circumscribed area of the image during repeated visits within the entire search of the image had to reach at least 1000 ms to be fed back. A 1000 ms dwell is considered to be a significant shift in visual attention when something of interest is detected. Viewing distance was 38 cm. In this study, .66 cm was equal to 1 deg on the mammographic display. Display size for single breast image was 18.4 x 14.5 cm, so that a circle with a 2.5 deg radius is 1.65/18.4= 9% zone of uncertainity in x, and 1.65/14.5= 11% zone of uncertainity in y in determining true hit location. The average size of a lesion target was 1 cm or 1.5 deg.

Results

The results of the Initial Overall-Impression phase are shown in Table 1.

Table 1. ROC for Initial Overall Impression

	ROC Area (Az)	
	Session 1	Session 2
R1	.672	.747
R2	.709	.622
R3	.754	.706
R4	.825	.755
R5	.588	.723
R6	.813	.724
Mean	.727	.713

Table 1 indicates small differences between readers' ROC area, Az performance, from Sessions 1 to 2 (F (1,5)= 0.13, n.s.). This suggests that practice effects were minimal

when feedback and control trials were counterbalanced across sessions. Readers overcalled positives in the initial overall-impression phase.

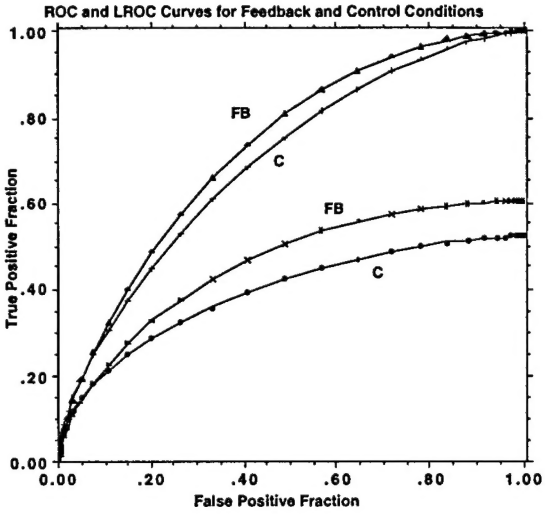
The results of the Final Decision Phase are shown in Table 2.

Table 2
ROC Area, LROC Area & Accuracy for Final Decision

	ROC Area		LROC Accuracy		LROC Area	
	Control	FB	Control	FB	Control	FB
R1	.686	.758	.468	.687	.373	.516
R2	.728	.728	.564	.595	.457	.457
R3	.720	.770	.632	.622	.439	.539
R4	.789	.751	.722	.673	.578	.502
R5	.541	.701	.140	.622	.082	.402
R6	.809	.719	.750	.513	.618	.437
$\overline{X}$=	.712	.738	.546	.619	.425	.476

Table 2 shows that mean ROC area, Az, and both mean LROC accuracy and mean LROC area increased from control to feedback. However, these differences were small and non-significant. The differences between control and feedback conditions were, for ROC area ($F(1,5) = 0.51$, n.s.), for LROC accuracy ($F(1,5) = 0.51$, n.s.) and for LROC area ($F(1,5) = 0.50$, n.s.) when tested using a repeated-measures ANOVA. The lower performance for LROC compared to ROC was due primarily to localization of wrong lesions.

Figure 1 shows the ROC curves for feedback (FB) and control (C) conditions for the 6 residents.



Discussion

In 1990, we found that mean area under the AFROC curve which also uses location in scoring performance, A1, improved from A1= .504 without CAVS feedback to A1= .618 with feedback in a pulmonary nodule-detection task using 6 radiology-resident readers (Kundel, Nodine, Krupinski, 1990). This was a 23% gain in performance as measured by area under the AFROC curve. In the present study we found that mean area under the LROC curve, which is a comparable measure, improved from .425 without CAVS feedback to .476 with feedback, only a 12% gain in performance. Both studies used similar experimental designs, but different eye-tracking systems. In addition to tracking eye position, the new system also monitored head movement. This was critical for the study of mammography because dynamic head movements occur during the reading of mammograms that involve searching for both masses and micro-calcifications requiring different viewing distances. The performance of only 3 of 6 readers improved with CAVS feedback, and the gains were small. When readers were questioned after the experiment, all 6 indicated that feedback merely confirmed their own perceptions about where they looked, but it did not help them decide whether or not a lesion was present in a feedback circle. Also, all readers indicated that they did not remember the cases from first to second sessions.

Our study of pulmonary nodule detection in 1990 found that the median cumulative gaze duration, varied as a function of decision outcome. Table 3 shows this.

Table 3. Comparison of Median Cumulative Gaze Duration (ms) for Decision Outcomes

	Pulmonary-Lesion Detection		Breast-Lesion Detection	
	Dwell	Rank	Dwell n	Rank
True Positive	2291	1	1370 (157)	2
False Negative	1283	3	860 (198)	3
False Positive	2091	2	2570 (95)	1
True Negative	547	4	400 (3018)	4
Wrong Lesion ------			1270 (119)	

The rank ordering of median gaze durations for the pulmonary-nodule detection task is similar to that for breast-lesion detection task. However, overall gaze durations are shorter, and the rank order of FPs and TPs is inverted in the breast-lesion task. It is difficult to identify the cause of this difference. The two tasks are different in a number of ways. The viewing session was limited to 15 sec for the pulmonary nodule study, and unlimited for present study. The readers were instructed to search for lesions in both tasks, but the test cases in the mammography task were deliberately chosen to contain subtle breast lesions in order to give room for improvement in performance if feedback was effective. But, this also probably made test-set difficulty greater than that for the pulmonary nodule-detection task. The consequence of this was that the subtle breast lesions attracted less attention and therefore fewer long dwell eye-fixation clusters.

Follow-Up Experiment

Because CAVS feedback was shown to be ineffective as a training tool for residents in the experiment reported above, we decided to test 3 experienced mammographers to see if feedback would aid them. Mammographers have more reading experience and therefore their search patterns should focus more often on true lesions giving CAVS feedback a better chance to work. We used the same set of test cases, but only gave the

Table 4

ROC Area, LROC Area & Accuracy for Final Decision for 3 Experienced Mammographers

	ROC Area		LROC Area	
	Control	FB	Control	FB
M1	.875	.939	.749	.877
M2	.744	.756	.488	.513
M3	.943	.722	.886	.444
Mean	.854	.806	.708	.611

mammographers one trial and limited the feedback to half of the 20 abnormal and normal cases. The results are shown in Table 4.

As Table 4 shows, CAVS feedback did not aid either ROC area, Az, performance (t=.542, df=4, n.s) or LROC area performance (t= .541, df=4, n.s.) for the experienced mammographers. However, CAVS feedback identified 70 of 136 (51%) possible breast lesions which is comparable to that found when CAVS feedback was applied to pulmonary nodules (53%), and only 12 out of 136 (9%) potential breast lesions failed to be fixated which is far less than 117 out of 480 (24%) found for residents in the CAVS pulmonary-nodule detection study. The differences between experienced mammographers and residents reading mammograms was reflected in both the fact that true lesions attracted attention of mammographers to a greater degree than residents. Despite this, however, CAVS feedback did not even aid the experienced mammographers.

Why Wasn't CAVS Feedback Effective as a Mammography Training Tool?
One can only speculate on the answer to this question.

First, an importance difference between using CAVS feedback to aid the detection of pulmonary nodules and using it to aid the detection of breast lesions was that the CAVS feedback presentation was randomized rather than blocked as during the presentation of the test cases for testing CAVS with breast lesions. This change in experimental protocol may not have given the readers enough of an opportunity to get "set" to deal with using a decision aid like visual feedback. In the pulmonary-nodule study, we blocked trials with and without feedback (10 cases per block) so that readers were able to "set" themselves in terms of readiness to interpret images using CAVS feedback. This may have led them to know when, and how accurately, CAVS feedback was working to track eye fixations.

Second, dynamic head movements of readers in the present experiment limited the accuracy of eye position data and this resulted in some misleading feedback. This may have influenced reader confidence in CAVS feedback as a decision aid. The problem of dynamic head movements was not present when we tested CAVS with pulmonary nodules because the readers head was constrained by a chin-and-head rest. This not only prevented dynamic head movements, but also guaranteed that readers viewed the display from a fixed distance. When CAVS was used with breast lesions, the viewing distance varied greatly because readers were searching for both breast masses and microcalcifications. The former are better seen from a distance, whereas the latter are seen better close up and so readers tended to move their head back and forth depending on which type of lesion was commanding attention.

Third, perhaps the most important reason why CAVS did not work has to do with limitations on the human visual system's object-recognition sensitivity. In general, except for perception of common everyday objects that are part of the visual world, most human object-recognition skills for man-made or machine-made visual images depend on perceptual learning. Snowden, Davies and Roling (2000) showed that sensitivity of

untrained adults' perceptual systems to mammographic targets like low-contrast dots and actual microcalcification clusters on mammograms improved significantly with perceptual learning. This reinforces our findings that amount of case-reading experience is directly related to performance accuracy in interpreting mammograms (Nodine, Mello-Thoms, 2000). However, despite the important role of perceptual learning in mammography, the call-back rate of experienced mammographers to resolve ambiguous image objects ranges from 5-10% for screening mammograms, and typically less than 1% of these ambiguous objects is verified as a true cancer. This suggest that, even among experts, image perception alone cannot resolve all potentially-abnormal objects detected during visual search of mammograms. Supplementary non-visual testing (e.g. ultrasound, biopsy) is required when visual scrutiny cannot clarify image perception.

Finally, our research, which focused on CAVS feedback to point out where attention was directed and concentrated during visual search may have been useful to the residents, but one feedback exposure per mammogram case could not substitute for multiple exposures to abnormal and normal objects encountered by practice reading mammogram cases with a tutor. Practice, which means reading mammograms and generating a plausible diagnostic interpretation, still seems to be the key for improving diagnostic performance of inexperienced readers, and our feedback practice was just too limited to benefit performance. CAVS feedback was also ineffective for experienced mammographers, but perhaps for a different reason, namely, that they had reached optimal performance using visual analysis alone.

When we tested CAVS on experienced mammographers in the follow-up study, it became clear that the cases that we selected for the test set contained very subtle cancers that pushed the object-recognition limits of even our experienced readers. They found that many of the cancer objects could not be resolved perceptually, and these cancers would probably fall into the "call-back" category in a clinical setting. Thus, without benefit of supplemental call-back information, the CAVS was ineffective even for these experienced mammographers.

Using CAVS with an Image-Analysis Decision Aid

We have been working on the use of an image-analysis decision aid in conjunction with CAVS. The image-analysis decision aid would offer feedback to the readers by analyzing the areas in the mammogram where the reader had significant visual dwell (>1000 ms) using wavelet analysis of the image. We create a compound image-information template by combining the spatial-frequency profiles of the areas that attracted prolonged visual dwell with the areas where the reader indicated the presence of a malignant lesion and running this through a filter bank. It is then possible to train a pattern classifier to map this compound image-information template into a like outcome based on image characteristics which we can feedback to the reader together with eye-position data. Using this as a supplemental decision aid to CAVS, the reader can gain information not only about what areas received prolonged visual attention during visual search, but also what is the likelihood that the reader's decision about the region of interest matches the decision outcome predicted by the pattern classifier. The decision aid is still in the development

stage and a preliminary report can be found under **Work in Progress, Study 2.** This paper will be presented at the Medical Image Perception Conference in September, 2001.

References

1. Kundel HL, Nodine CF, Krupinski EA. Computer-displayed eye position as a visual aid to pulmonary nodule detection. Invest Radiol 1990; 25: 890-896.

2. Krupinski EA, Nodine CF, Kundel HL. Enhancing recognition of lesions in radiographic images using perceptual feedback. Opt Eng 1998; 37: 813-818.

3. Nodine CF, Kundel HL, Mello-Thoms et al., How experience and training influence mammography expertise. Acad Radiol. 1999;6:575-585.

4. Nodine CF, Mello-Thoms C. The nature of expertise in radiology. In J. Beutel, HL Kundel, R L VanMetter (Eds.) Handbook of medical imaging. Vol. 1 Physics and psychophysics. Bellingham, WA, SPIE Press; 2000:859-894.

5. Swensson RG. Using localization data from image interpretations to improve estimates of performance accuracy. Med Decision Making 2000; 20: 170-185.

6. Krupinski EA, Nodine, CF. Gaze duration predicts the location of missed lesions in mammography. In A. G. Gale et al., editor, Digital Mammography, Elsevier Science B.V., 1994.

TASK 7: Analysis of CAVS Experiment , COMPLETE.

(6.6) Work in Progress

We are currently working on two studies that we will be reporting at the Medical Image Perception Conference IX on September 20-23, 2001.

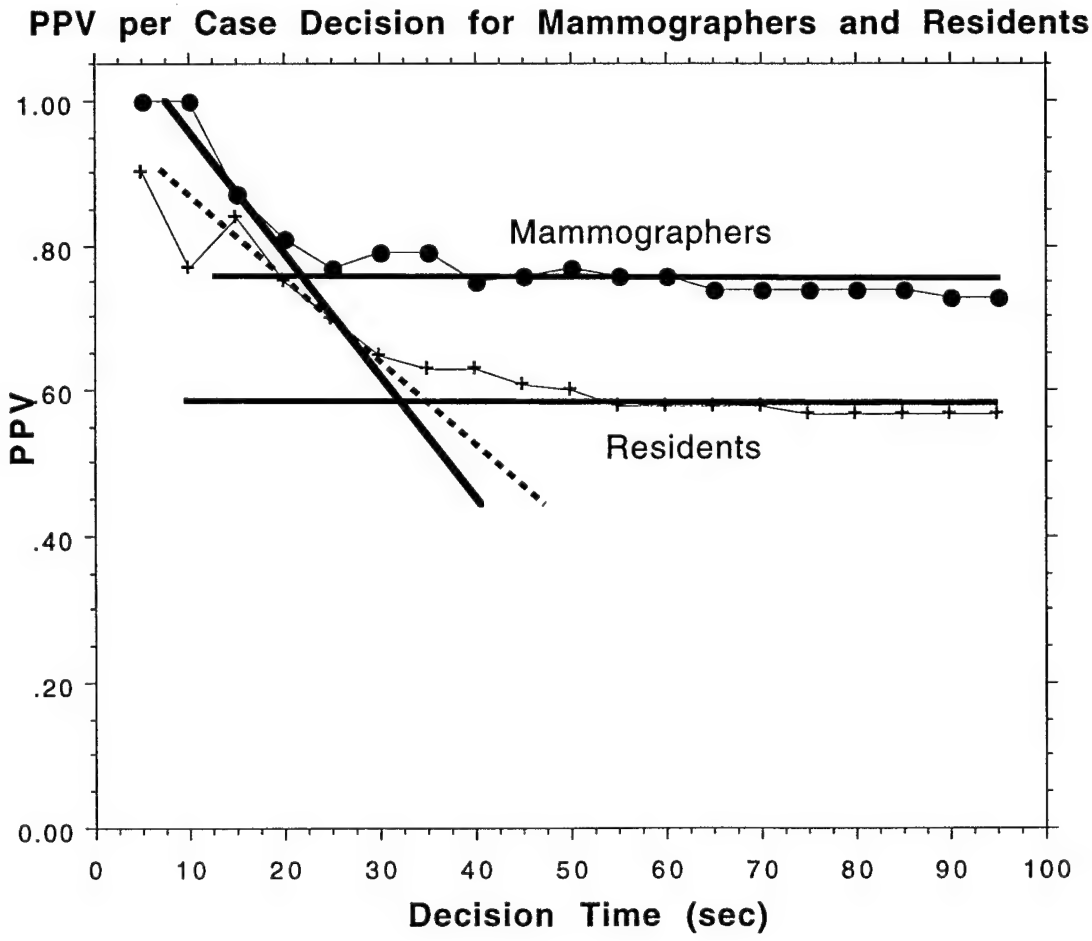
Study 1: Timing the Course of Reader Decisions during Mammographic Interpretation. Calvin F. Nodine, Claudia Mello-Thoms, Harold L. Kundel Pendergrass Lab, University of Pennsylvania

Twenty years ago Christensen et al (1981) studied what they called "time-perception curves" of experienced and inexperienced radiologists detecting abnormalities in chest radiographs. Their analysis of the time-perception curves suggested that the medical image interpretation consists of two components of perception, one rapid and the other slow. Christensen et al. drew heavily on a prior study done in our lab which showed that radiologists could detect many abnormalities in chest radiographs within a 200 ms presentation, and that performance improved when given additional viewing time to focally search the image (Kundel & Nodine, 1975). This led us to propose that image perception during visual search consists of a "global response" (the fast component) and "checking fixations" (the slow component). We have elaborated on this model of visual search since then, and the present study examines time perception curves in two studies in which radiologists search mammograms for breast abnormalities.

We examined the time spent searching local areas of mammograms for breast abnormalities with eye fixations by measuring positive predictive value (ppv) which is: true positive decisions/ true-positive + false-positive decisions as a function of decision time which is less complicated than that used by Christensen et al. to study time-perception curves. We measured eye position in the two studies involving a total of 10 readers of different levels of experience (<1000 case readings per year vs. >4000 case-readings per year) searching for breast abnormalities in test sets of 40 mammograms.

Using ppv vs. decision time , we show how decision performance is related to reader experience over the time course of image interpretation. We also show the time course of eye-fixation dwell times related to these decision outcomes. For example, most long dwells occur early in viewing. Finally, compare the time courses of readers with different levels of mammography case-reading experience.

Figure 2 compares ppv (positive predictive value) vs. decision time for mammographers and residents and shows the initial "fast" component (negative slope) and subsequent "slow" component (flat slope) as the model predicts with experienced mammographers performing significantly better than residents.



Preliminary results suggest that the two-component model generalizes to the detection of breast abnormalities in mammograms. The global response initiates search by giving the reader a Gestalt overview which flags conspicious breast abnormalities that are skimmed off with minimal scanning search. The global response then directs the slow component, scanning search, in which checking fixations flesh out both conspicuous and more subtle, breast abnormalities by visually scrutinizing image features. The two-component search model suggests that the reader should initially scan the image to obtain an overall impression, e.g. Normal or Abnormal, by taking advantage of the global response, and then focus scanning search using checking fixations on areas of the mammogram that were "flagged" as perturbed from what the reader would expect to see in a normal case.
It is important that the reader continue searching perturbed areas until satisfied that all perturbations have received attention. Christensen et al. found that experienced readers terminate search of chest radiographs at a point in the time course of decision making when they are still making more true than false positive decisions. Our analysis of the time course of searching mammograms for breast abnormalities indicates that both residents and mammographers terminate search when true positives equal false positives. However, the level of performance as measured by ppv is about 20% higher for mammographers than residents.

Study 2: On the Image-Based Nature of Decision Outcome in Mammogram Interpretation. Claudia Mello-Thoms, Calvin F. Nodine and Harold L. Kundel
University of Pennsylvania School of Medicine

The goal of screening mammography is the early detection of breast cancer. To that end image presentation and analysis algorithms are developed. Nonetheless, an aspect that is usually overlooked is the radiologist's ability to correctly interpret the information being displayed. In this way, it is important to understand the underlying decision making mechanisms that lead radiologists either to report, or miss, a malignant lesion in the breast, as well as to mistake normal tissue as being malignant.

In this study we investigated the decision outcomes of experienced mammographers and less experienced radiology residents when reading mammogram cases in search of malignancy. We analyzed the areas of the mammogram that elicited a response by the observers (whether this response is correct or not), as well as lesion free areas that attracted prolonged visual dwell (>1000 ms) but where no response was made, and lesion-containing areas that went unreported. These areas were segmented from the mammogram and analyzed using a filter bank based on wavelet analysis that zoomed-in on high-frequency phenomena, as well as had good spatial frequency localization of low-frequency phenomena. This analysis allowed us to model, in the spatial frequency domain, the visually-selected image areas. The differences in the spatial frequency representation of the image areas that lead to the four decision outcomes (True and False Positives, True and False Negatives) were then compared using analysis of variance, and post-hoc tests were conducted to determine the statistical significance of the differences.

We have shown in a preliminary study that spatial frequency differences from wavelet analysis of the mammograms vary among the four decision outcomes, and, that some of

these differences were statistically distinctive for different decision outcomes (Mello-Thoms et al, 2001). Furthermore, we have shown that these differences can be used by a artificial neural network (ANN) pattern classifier to predict how a given observer will respond to the image elements present in a new mammogram, based upon the observer's past responses to different cases. In this context the ANN performed a mapping between a 14-dimensional input space (the energy values of the wavelet packets decomposition of the area of the mammogram under consideration in different spatial frequency bands) and a 1-dimensional output space (the predicted decision outcome). Using this mapping, our system yielded the following rates of correct predictions

Decision Outcome	Percent of Correct Prediction
TP	64
FP	77
TN	70
FN	28

Note that the prediction rates for the False Negative decision outcomes was not good, at only 28%. One of the main reasons for that was the limited number of such samples in the data set, which made it very difficult for the ANN to create an appropriate internal representation of this type of decision outcome.

In the present study we attempted to determine an individual profile, for each observer, in terms of that observer's hits and misses. This profile was correlated with the observer's experience level (as measured by the number of cases read by the observer over a given period of time, which has been shown to positively correlate with performance (Nodine et al, 1999)). The aim of such profiling was to determine common errors made by less experienced observers, as well as how such errors differ from correct decisions (for example, how True Positives compare with both False Positives and with False Negatives).

Reference:

Christensen et al. The effect of search time on perception. Radiol 1981;138:361-365.
Kundel & Nodine Interpreting chest radiographs without visual search Radiol 1975:527

C. Mello-Thoms, S.M. Dunn, C.F. Nodine, H.L. Kundel. An analysis of perceptual errors in reading mammograms using quasi-local spatial frequency spectra. Journal of Digital Imaging, 2001 (in press).

C.F. Nodine, H.L. Kundel, C. Mello-Thoms, et al. How experience and training influence mammography expertise. Academic Radiology 1999; 6:575-585.

(7) KEY RESEARCH ACCOMPLISHMENTS:

Our research studies in 2000-2001 have led to three key findings:

1. CAVS feedback was tested as a mammography training tool and the results on 6 radiology residents and 3 experienced mammographers indicate that visual feedback which simply points out locations on the mammogram where readers focus attention was not an effective training tool leading for improving the detection of breast cancers. However, using CAVS feedback in conjunction with an image-analysis decision aid currently under development has potential for making the training tool more effective.

2. A two-component model of image perception was tested and compared for detecting cancers in chest radiograms and cancers in mammograms. A plot of performance as measured by predictive value of a positive decision as a function of decision time supports a two-component model and demonstrates significant performance differences as a function of case-reading experience.

3. Error in reading mammograms as measured by Positive Predictive Value (true-positive decisions/ true-positive + false-positive decisions) increases with decision time for both experienced mammographers and radiology residents. The optimal trade-off for maximum performance in reading mammograms was found to be 30 sec. in our study.

(8) REPORTABLE OUTCOMES:

In addition to the work completed and in progress as discussed above, we have completed the following articles:

1. Nodine CF, Kundel HL, Mello-Thoms C and Weinstein SP. Role of computer-assisted visual search in mammographic interpretation. Proc SPIE: Image Perception and Performance 2001; 4324: 52-55. (Appendix 1).

2. Mello-Thoms C, Dunn SM, Nodine CF, Kundel HL. Using computer-assisted perception to determine the characteristics of missed and reported breast cancers. Proc SPIE: Image Perception and Performance 2001; 4324: 64-67. (Appendix 2).

3. Nodine CF, Mello-Thoms C, Weinstein SP et al. Do subtle breast cancers attract visual attention during initial impression? Proc SPIE: Image Perception and Performance 2000; 3981; 156-159. (Appendix 3).

4. Mello-Thoms C, Dunn SM, Nodine CF, Kundel HL. Image structure and perceptual errors in mammogram reading. Proc SPIE: Image Perception and Performance 2000; 3981:170-173. (Appendix 4).

5. Mello-Thoms, Nodine, Weinstein et al. An Unobtrusive Method for Monitoring Visual Attention During Mammogram Reading" Proc SPIE: Image Perception and Performance 2000; 3981:160-163. (Appendix 5).

6. Nodine CF, Mello-Thoms C, Weinstein SP et al. Blinded review of retrospectively visible but unreported breast cancers: An eye position analysis. Radiology 2001; 221: October in press.

7. Mello-Thoms C, Dunn SM, Nodine CF, Kundel HL. An analysis of perceptual errors in reading mammograms using quasi-local spatial frequency spectra. Journal of Digital Imaging 2001, in press.

We are currently working on four papers:

1. " Timing the Course of Reader Decisions during Mammographic Interpretation." Calvin F. Nodine, Claudia Mello-Thoms, Harold L. Kundel, 2001, in preparation.

2. "On the Image-Based Nature of Decision Outcome in Mammogram Interpretation." Claudia Mello-Thoms, Calvin F. Nodine and Harold L. Kundel, 2001, in preparation.

3. "An analysis of perceptual errors in reading mammograms using quasi-local image frequency spectra." Mello-Thoms, Dunn, Nodine CF et al., 2001, in preparation.

4. "Determining the quasi-local spatial frequency characteristics of missed and reported breast cancers." Mello-Thoms, Dunn, Nodine, Kundel, Weinstein et al., 2001, in preparation.

(9) CONCLUSIONS

Goals

The primary goal of the project is to develop a mammography training tool to improve perceptual and cognitive skills of observers leading to mammographic expertise.

Prerequisites to this goal were an understanding of: (a) how mammographers are trained, (b) what skills are required to carry out the task of detecting, classifying and diagnosing abnormalities in mammograms, and (c) the effectiveness of current mammography training measured by evaluating the performance of residents using a test-set of mammograms representing various abnormalities. We have examined these three questions and reported the results in three articles (Nodine, Kundel, Mello-Thoms, 1999; Nodine, Mello-Thoms, 2000; Nodine, Mello-Thoms, Weinstein et al., 2001, in press).

When we come to the question of how can perceptual and decision-making skills be improved? The answer that our research seems to be saying is: "Practice Makes Perfect". This is a deceptively simple answer.

Importance of Computer-Training Tools for Mammography

During their medical training, radiologists have to learn much more than simply how to read mammograms, and there is not enough training time in the radiology residency program to make experts in mammography because mammography is but one of many radiology specialities. Rather, what may be needed is a more effective way to train residents during their clinical residency in mammography. We need to supplement apprenticeship mentoring by expert computer-training tools. Expert computer tools like CAVS can provide systematic feedback tailored specifically to each resident's level of training and experience. However, a formal test of CAVS feedback that pointed out regions of the mammogram deemed "suspicious" based on analysis of eye-position dwell times indicated that this perceptual aid alone was shown to be ineffective as a training tool for improving performance of either radiology residents or experienced mammographers.

Evaluating CAVS Feedback as a Training Tool for Mammography

We initially proposed CAVS feedback as a training tool for mammography because it had been effective in improving pulmonary-nodule detection performance of radiology residents. To determine why CAVS feedback did not work with breast lesions but did work with pulmonary nodules in a prior study (Kundel, Nodine, Krupinski, 1990) we compared the efficiency of CAVS in the two studies. In the current breast-lesion test, CAVS feedback correctly localized only 87 out of 272 (32%) possible breast lesions which is considerably less than 254 out of 480 (53%) found when CAVS feedback was tested with pulmonary nodules. We calculated the probability of a CAVS feedback localizing a breast lesion by chance as 21%. In the CAVS study with residents as readers, 87 out of 272 (32%) possible true lesions were localized. This was better than chance, but considerable less than the 53% found when CAVS was tested with pulmonary lesions. In addition, 96 out of 272 (35%) possible true lesions were not fixated at all. This is 11% more complete misses than 117 out of 480 (24%) found when CAVS was tested with pulmonary nodules. These differences suggest that reading mammograms is different from reading chest radiographs and may be a more difficult task for radiology residents in training.

The experienced mammographers were closer in agreement to the data found in the pulmonary nodule study. CAVS feedback efficiency of mammographers was 70 out of 136 (51%) true breast lesions localized, and only 12 out of 136 (9%) breast lesions that failed to be fixated. Despite this, CAVS did not improve the performance of experienced mammographers, perhaps because they had reached asymptotic performance within the constraints of the reading task which was imposed upon them of interpreting two-views of a mammogram image by depending solely on visual analysis.

Why didn't CAVS Work in Mammography?

Our findings after testing CAVS as a training tool for mammography seem to indicate that radiology resident readers could not take advantage of CAVS feedback because they did not fixate the true lesions as part of their evaluation of the mammogram cases. This is supported by low LROC area performance shown in Table 2. Basically, radiology residents could not reliably distinguish true lesions from false lesions in the mammogram

test set, and, CAVS feedback cannot work unless the true lesions are fixated. In contrast, the experienced mammographers did fixate true lesions as evidenced both by increased CAVS efficiency (19% increase) and a significant increase in lesion-detection accuracy as indicated by LROC area performance in Table 4. However, the breast lesions that the mammographers missed may have been too difficult to detect and resolve purely by visual analysis. Mammograpers often supplement visual analysis by non-visual tests (e.g. ultrasound, biopsy) which were not available. Even if CAVS feedback did correctly localize the difficult true lesions, the mammographers were unable, on the basis of image-feature analysis, to decide whether or not a true lesion was being localized. This conclusion suggests that CAVS needs to be supplemented by additional decision aids that analyze the image content contained within the CAVS feedback localization areas. Thus, even though CAVS feedback was found to be ineffective as a teaching tool, it did provide a method of sampling the mammographic image which may prove to be valuable for directing decision aids designed to analyze reader-determined perceptually-interesting areas of the mammographic image. We are beginning to work on this problem.

(10) REFERENCES

1. Krupinski EA, Nodine CF, Kundel HL. Enhancing recognition of lesions in radiographic images using perceptual feedback. Opt Eng 1998;37: 813-818.

2. Nodine CF, Kundel HL, Mello-Thoms et al., How experience and training influence mammography expertise. Acad Radiol. 1999;6:575-585.

3. Kundel Hl, Nodine CF, Krupinski EA. Computer-displayed eye position as a visual aid to pulmonary nodule interpretation. Invest Radiol. 1990; 25: 890-896.

4. Krupinski EA. Visual scanning patterns of radiologists searching mammograms. Acad Radiol. 1996; 3: 137-144.

5. Nodine CF, Mello-Thoms C, Weinstein SP et al. Blinded review of retrospectively visible but unreported breast cancers: An eye position analysis. Radiology 2001; 221: October in press.

(11) APPENDICES

1. Nodine CF, Kundel HL, Mello-Thoms C and Weinstein SP. Role of computer-assisted visual search in mammographic interpretation. Proc SPIE: Image Perception and Performance 2001; 4324: 52-55. (Appendix 1).

2. Mello-Thoms C, Dunn SM, Nodine CF, Kundel HL. Using computer-assisted perception to determine the characteristics of missed and reported breast cancers. Proc SPIE: Image Perception and Performance 2001; 4324: 64-67. (Appendix 2).

3. Nodine CF, Mello-Thoms C, Weinstein SP et al. Do subtle breast cancers attract visual attention during initial impression? Proc SPIE: Image Perceptionand Performance 2000; 3981; 156-159. (Appendix 3).

4. Mello-Thoms C, Dunn SM, Nodine CF, Kundel HL. Image structure and perceptual errors in mammogram reading. Proc SPIE: Image Perception and Performance 2000; 3981:170-173. (Appendix 4).

5. Mello-Thoms, Nodine, Weinstein et al. An Unobtrusive Method for Monitoring Visual Attention During Mammogram Reading" Proc SPIE: Image Perception and Performance 2000; 3981:160-163. (Appendix 5).

Appendix 1

Role of Computer-Assisted Visual Search in Mammographic Interpretation.

Calvin F. Nodine, Harold L. Kundel, Claudia Mello-Thoms and Susan P. Weinstein
University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6086

Abstract

We used eye-position data to develop Computer-Assisted Visual Search (CAVS) as an aid to mammographic interpretation. CAVS feeds back regions of interest that receive prolonged visual dwell (≥ 1000 ms) by highlighting them on the mammogram. These regions are then re-evaluated for possible missed breast cancers. Six radiology residents and fellows interpreted a test set of 40 mammograms twice, once with CAVS feedback (FB), and once without CAVS FB in a crossover, repeated-measures design. Eye position was monitored. LROC performance (area) was compared with and without CAVS FB. Detection and localization of malignant lesions improved 12% with CAVS FB. This was not significant. The test set contained subtle malignant lesions. 65% (176/272) of true lesions were fixated. Of those fixated, 49% (87/176) received prolonged attention resulting in CAVS FB, and 54% (47/87) of FBs resulted in TPs. Test-set difficulty and the lack of reading experience of the readers may have contributed to the relatively low overall performance, and, may have also limited the effectiveness of CAVS FB which could only play a role in localizing potential lesions if the reader fixated and dwelled on them.

Keywords: Breast Lesions, Computer Aids, Visual Search, Visual Feedback, Eye Fixations, LROC

1.Introduction

Ten years ago we published a study showing that when radiology residents re-examined chest x-ray images with visual feedback, nodule detection performance was significantly higher than when they re-examined the same chest images without visual feedback (Kundel, Nodine, Krupinski, 1990). In this study, we found a significant 23% improvement in AFROC performance with visual feedback.

The present paper will report a study showing that visual feedback did not lead to a significant improvement in lesion-detection performance for a group of radiology residents and radiology fellows searching mammograms for breast cancers. Our current study shows only a 12% improvement in LROC performance (area) detecting malignant breast lesions. Both studies used similar experimental designs, but different eye-tracking systems. In addition to tracking eye position, the new system also monitored head movement. This is critical for the study of mammography because dynamic head movements occur during the reading of mammograms that involve searching for both masses and micro-calcifications requiring different viewing distances.

The rationale for feeding back regions of prolonged attention, indicated by eye-fixation clusters, is that, we have shown that false negative dwell times are more than twice as long in cumulative gaze duration as true negative dwell times (Kundel. Nodine, Krupinski. 1990). Cumulative gaze duration is defined by the sum of fixation durations within a cluster of fixations. Thus, feeding back regions on a chest or breast image where attention is concentrated, as indicated by cumulative gaze duration (dwell for short), but no decision is reported, may help identify fixated but unrecognized lesions. We have used 1000 ms as the dwell threshold for identifying feedback (FB) regions because it maximizes the likelihood of feeding back lesion-containing areas without feeding back numerous lesion-free areas (Krupinski, Nodine, Kundel, 1998). Because visual feedback may also prompt false-positive decisions, it is necessary to evaluate the effectiveness of visual feedback using a performance measure that adjusts for response bias. Both AFROC and LROC methods are ideal for this purpose and in addition, LROC gives us a measure of lesion-localization accuracy by differentiating between correctly localized TPs and incorrectly localized lesions on lesion-containing cases, called wrong lesions (WLs) (Swensson, 2000).

2.Materials and Methods

A crossover, repeated measures design was used in which 6 readers were assigned to experimental conditions by counterbalancing testing order and by randomizing the sequencing of 40 feedback and 40 non-feedback trials over two sessions separated by one month in order to allow for forgetting. The 40 mammogram test set was presented on a high resolution 21" digital workstation (Clinton Electonics DS 5000L, Rockford, Il, 2560 x 2048). Each case was displayed in two views, craniocaudal (CC) in the left half of the screen and mediolateral oblique (MLO) in the right half of the screen.

The test set of 40 mammogram cases, half containing malignant lesions and half lesion free, was presented to 6 readers, three radiology fellows and 3 radiology residents, having limited case reading experience (range 302-976 case readings prior to testing). A case consisted of two views of a single breast. All malignant-lesion cases but one had a single lesion in each view. One case had four lesions in each view. There were 13 cases with masses, 6 cases with microcalcifications, 1 case with architectural distortion and 20 lesion-free normal cases. The cases were selected by an experienced mammographer and represented "subtle lesion" mammogram cases. Each case was digitized using a 50 micron spot size by a Lumiscan 100 digitizer (Lumysis, Sunnyvale, CA).

Each trial consisted of an initial overall-impression phase and final-decision phase. During the overall impression, readers were asked to evaluate CC and MLO views of each digitized breast case and decide whether or not it contained a malignant lesion. Eye-position was recorded using an ASL 4000 SU eye-head tracker (Applied Science Labs, Bedford, MA). When readers indicated they had finished evaluating each case, eye-position recording was terminated, and a menu was displayed. Readers used a mouse cursor to mark their overall-impression as either normal or abnormal and their initial decision confidence: high; medium; or, low. The final-decision phase followed during which eye-position data were analyzed and feedback regions calculated (regardless of whether feedback or control trial). During final decision the reader was asked either to re-evaluate the FB regions, or the entire image. If a malignant lesion was detected either from the overall impression, or newly discovered during the final decision, the reader clicked on the lesion. This called up a menu and readers indicated lesion type: mass; microcalifications; or, architectural distortion, and gave final decision confidence of malignancy of the lesion: high; medium; or, low.

Readers were told that they could change their mind from initial overall-impression phase to final-decision phase and it was stressed that they should respond "abnormal" only if they consider that the case probably contained an malignant lesion. During the final-decision phase the reader was asked to localize malignant lesions in CC and MLO views (if possible) on cases called "abnormal". However, it was emphasized that this did not preclude localization of newly discovered lesions during the final-decision phase on cases called "normal" initially. Conversely, readers could decide during final-decision phase that an initial "abnormal" decision was in error, and decide the case was free of a malignant lesion (i.e. "normal").

The decision data for the initial overall-impression phase and final-decision phase were analyzed separately. In the overall impression, reader confidences were used to construct a 2 x 6 ROC truth table. In the final decision, LROC scoring was applied to determine how many localized lesions matched true lesion locations as determined by the mammographer. If a localized lesion fell within 2.5 deg (1.65 cm) of a true lesion location, and was given the highest confidence for the case, it was scored as a true positive (TP). If a localized lesion was outside the 2.5 deg zone of a true lesion, and was given the highest confidence for the case, it was scored as a wrong lesion (WL). In tie cases the TP won (Rule 2). Lesion-free cases were scored as true negative (TN) if no lesions were localized, but if lesions were localized, the lesion with the highest confidence rating for the case was assigned as a false positive (FP). The scoring of decision outcomes by case with associated confidence ratings were used to construct a 3 x 6 LROC truth table. Hits for decision outcomes in the truth table were defined as follows: Hit zone= 2.5 deg, Minimum dwell for CAVS FB region= $\geq$ 1000 ms. Viewing distance = 38 cm. In this study, .66 cm = 1 deg on the mammographic display. Display size for single breast image was 18.4 x 14.5 cm, so 1.65/18.4= 9% zone of uncertainity in x, and 1.65/14.5= 11% zone of uncertainity in y in determining true hit location. The average size of a lesion target was 1 cm or 1.5 deg.

3.Results

The results of the initial overall-impression phase were analyzed by ROC. The results are shown in Table 1.

Table 1. ROC for Initial Overall Impression

	ROC Area (Az)	
	Session 1	Session 2
R1	.672	.747
R2	.709	.622
R3	.754	.706
R4	.825	.755
R5	.588	.723
R6	.813	.724
Mean	.727	.713

Table 1 indicates small differences between readers' Az performance from Sessions 1 to 2 ($F(1,5)= 0.13$, n.s.) , so this suggests that practice effects were minimal when feedback and control trials were counterbalanced across sessions. Readers overcalled positives in the initial overall-impression phase.

The results of the Final Decision are shown in Table 2.

Table 2
ROC Area, LROC Area & Accuracy for Final Decision

	ROC Area		LROC Accuracy		LROC Area	
	Control	FB	Control	FB	Control	FB
R1	.686	.758	.468	.687	.373	.516
R2	.728	.728	.564	.595	.457	.457
R3	.720	.770	.632	.622	.439	.539
R4	.789	.751	.722	.673	.578	.502
R5	.541	.701	.140	.622	.082	.402
R6	.809	.719	.750	.513	.618	.437
$\bar{X}$=	.712	.738	.546	.619	.425	.476

Table 2 shows that mean ROC area, Az, and both mean LROC accuracy and mean LROC area increased from control to feedback. However, these differences were small and non-significant. The differences between control and feedback conditions were, for ROC area ($F(1,5)= 0.51$, n.s.), for LROC accuracy ($F(1,5)= 0.51$, n.s.) and for LROC area ($F(1,5)= 0.50$, n.s.) when tested using a repeated-measures ANOVA. The lower performance for LROC compared to ROC was due primarily to localization of wrong lesions.

4.Discussion

In 1990, we found that mean AFROC performance improved from A1= .504 without FB to A1= .618 with FB in a pulmonary nodule-detection task using 6 radiology-resident readers. A 23% gain. Today we found that mean LROC area, which is a comparable measure, improved from .425 without FB to .476 with feedback, only a 12% gain. The performance of only 3 of 6 readers improved with CAVS feedback, and the gains were small. When readers were questioned after the experiment, all 6 indicated that feedback merely confirmed their own perceptions about where they looked, but it did not help them decide whether or not a lesion was present in a FB circle. Also, all readers indicated that they did not remember the cases from first to second sessions.

Our study of pulmonary nodule detection in 1990 found that the median cumulative gaze duration, varied as a function of decision outcome. Table 3 shows this.

Table 3. Comparison of Median Cumulative Gaze Duration (ms) for Decision Outcomes

	Pulmonary-Lesion Detection		Breast-Lesion Detection (n)	
	Dwell	Rank	Dwell	Rank
True Positive	2291	1	1370 (157)	2
False Negative	1283	3	860 (198)	3
False Positive	2091	2	2570 (95)	1
True Negative	547	4	400 (3018)	4
Wrong Lesion	------		1270 (119)	

The rank ordering of median gaze durations for the pulmonary-nodule detection task is similar to that for breast-lesion detection task. However, overall gaze durations are shorter, and the rank order of FPs and TPs is inverted in the breast-lesion

task. It is difficult to identify the cause of this difference. The two tasks are different in a number of ways. The viewing session was limited to 15 sec for the pulmonary nodule study, and unlimited for present study. The readers were instructed to search for lesions in both tasks, but the test cases in the mammography task were deliberately chosen to contain subtle breast lesions in order to give room for improvement in performance if feedback was effective. But, this also probably made test-set difficulty greater than that for the pulmonary nodule-detection task. The consequence of this was that the subtle breast lesions attracted less attention and therefore fewer long eye-fixation clusters. However, this may not be the cause of the differences. Krupinski and Nodine (1994), and Krupinski, Nodine, Kundel (1998), found median gaze durations as a function of decision outcome for radiologists searching mammography cases closer to those found for the pulmonary-nodule detection task.

As a consequence of shorter median gaze durations as a function of decision outcome, only 49% of FNs (85/172) out of all lesion-containing regions had dwell greater than 1000 ms in the present experiment, whereas 59% of FNs had dwell ≥ 1000 ms in the pulmonary-nodule detection study. Only 10% (55/544) of all lesions were not looked at (fixated). Survival curves show differences as a function of decision outcome (Krupinski, Nodine, Kundel, 1998).

The effectiveness of feedback can be compared to that in the pulmonary nodule study. CAVS FB hit 46% (40/87) missed lesions in the present study which is close to 42% reported in the pulmonary nodule study. It takes 11- 7 deg FB circles to cover the average breast image, compared to 26- 5 deg circles to cover the average lung fields in a PA chest image. An average of 3.6 FB circles were generated per breast image, compared with an average of 5 FB circles per chest image. The probability of a FB circle hitting a breast lesion was 1/11 or 9%, compared to 1/26 or 4% in the pulmonary nodule study. Multiplying 3.6 FB circles for the average breast x 9%= 32% as chance hitting of a breast lesion, compared to 5 FB circles x 4%= 20% chance hitting a lung nodule. Therefore we can conclude that 87 TPs & FNs/272 lesions= 32% actual hit rate is equivalent to chance hitting of breast lesions, compared to better than chance hitting of pulmonary nodules.

The reading experience of the readers in the present experiment was low. The fellows had read about 1000 mammogram cases, and the residents had only read 300-500 cases prior to the experiment. The test-set difficulty and the lack of reading experience may have contributed to the relatively low overall performance, and may have also limited the effectiveness of CAVS feedback which could only play a role in localizing potential lesions if the reader fixated and dwelled on them. Inexperienced readers have seen relatively few true cancers and thus it is not surprising that they had difficulty differentiating true cancers from wrong lesions which often received higher decision confidence when both were localized on the same case. Why didn't FB work? 1. Very neutral instructions combined with random presentation of FB. 2. Dynamic head movement of readers limited accuracy of eye position data resulting in misleading FB. This may have influenced reader confidence in FB.

5.Acknowledgement

This work was partially supported by Grant DAMD17-97-1-7103 between the USAMRMC and C. F. Nodine.

6.Reference

1. Kundel HL, Nodine CF, Krupinski EA. Computer-displayed eye position as a visual aid to pulmonary nodule detection. Invest Radiol 1990; 25: 890-896.

2. Krupinski EA, Nodine CF, Kundel HL. Enhancing recognition of lesions in radiographic images using perceptual feedback. Opt Eng 1998; 37: 813-818.

3. Swensson RG. Using localization data from image interpretations to improve estimates of performance accuracy. Med Decision Making 2000; 20: 170-185.

4. Krupinski EA, Nodine, CF. Gaze duration predicts the location of missed lesions in mammography. In A. G. Gale et al., editor, Digital Mammography, Elsevier Science B.V., 1994.

Using Computer Assisted Perception to Determine the Characteristics of Missed and Reported Breast Cancers

Claudia Mello-Thoms①, Stanley Dunn②, Calvin F. Nodine①, Harold L. Kundel① and Susan P. Weinstein①

①University of Pennsylvania School of Medicine, Department of Radiology, Philadelphia PA 19104

②Rutgers University, Department of Biomedical Engineering, Piscataway NJ 08855-0909

Abstract

Early detection of breast cancer is the desired goal in breast cancer screening. Nonetheless it has been reported in the literature that 10-30% of all breast cancers are missed by the radiologist [1], albeit most of these are deemed visible in retrospect on the mammogram. In this work we have studied the underlying structure of the areas that attracted the radiologist's visual attention and either yield or do not yield a response.We have shown that the spatial frequency profile of areas where a lesion is detected (TP) is significantly different from the one where a lesion is missed (FN), where a lesion is incorrectly placed (FP) or of lesion-free areas that are correctly identified (TN). Furthermore, we have shown that the spatial frequency profile alone can be used by an artificial neural network to predict decision outcome in that area of the image.

Keywords: Image structure, perceptual errors, mammogram reading, wavelet packets.

1.Introduction

In the detection of breast cancer, catching the abnormality as early as possible can significantly change the prognosis for a woman diagnosed with this disease. Consequently, renewed efforts have been made to develop accurate imaging techniques that can display abnormalities of smaller sizes. Nonetheless, a problem that is usually overlooked when considering such imaging techniques is the radiologist's ability to correctly interpret what is on the image. It has been shown [1] that 10-30% of all breast cancers are missed, being only found retrospectively. From these, 65% are fixated by the high-resolution central/foveal vision [2]. Thus, these cancers are not missed because of search errors, but because of perception and decision-making errors. Additionally, there exists another cost to detect breast lesions while in their initial stages, namely, that of the number of incorrect decisions (False Positives) generated. It has been estimated that over a period of 10 years 1/3 of the women undergoing periodic mammographic screening will have a FP test result [3]. These FPs cause undue stress to women and significantly raise the cost of health care [3]. In this way, it is necessary to develop a tool capable of helping radiologists not only to miss fewer abnormalities but also to make fewer mistakes when detecting these abnormalities.

In this paper we will develop one of such tools. We will show that different decision outcomes (such as True and False Positives and Negatives) possess different signatures in the spatial frequency domain, and that these signatures can be used to uniquely identify the decision outcome. Futhermore, after learning how experienced radiologists respond to the elements present in a mammographic training set, we will use the spatial frequency signatures of areas in a mammographic test set, that attracted the radiologists' visual attention, to predict how the radiologists will respond to such elements. This system, called Computer Assisted Perception (CAP), predicts four types of decision outcome: True Positives (TPs), False Positives (FPs), True Negatives (TNs) and False Negatives (FNs).

2.Materials and Methods

Four experienced observers from the staff of the Hospital of the University of Pennsylvania (HUP) read 40 two-view (cranio-cauldal, CC, and medio-lateral-oblique, MLO) mammogram cases. Twenty cases contained abnormalities that were not reported in first viewing, being only found retrospectively. Ten cases contained abnormalities that were reported in first viewing, and ten cases were cancer free, being stable for two years. All cancer cases were biopsy proven. These cases were obtained from the archives of HUP by one of the authors, who is a mammographer but did not participate as an observer in this experiment (SPW). The films were digitized using a Lumiscan Model 100 digitizer (Lumisys Inc, Sunnyvale, CA), with a 50 microns spot size. The two-views were displayed on a single 21-inches landscape 2560 x 2048 gray scale monitor (ORWIN, Model DS5000L, Amityville, NY).

The observers were instructed to search for malignancy, and freely examined the cases until they felt confident to indicate, using a mouse-controlled cursor, if and where a malignant lesion was present. The eye position of the observers was monitored during search using an eye-tracker ASL4000SU (Applied Science Laboratories, Waltham, MA), and both eye-

positions and responses, indicated with the mouse-controlled cursor, were used to determine the areas in the image that attracted the observers' visual attention.

The regions that yield responses by the observers (True and False Positives), as well as lesion free regions that attracted visual dwell but did not elicit a response (True Negative response), and regions that contained a lesion that went unreported (False Negative response), were extracted from the images. Each region corresponded to a 5° visual angle, and each region was labeled by the type of decision outcome that it yielded. Namely, there were four types of regions, True Positives (an existing lesion was correctly reported by the observer), False Positive (a non-existing lesion was reported by the observer), False Negatives (an existing lesion was not reported by the observer) and True Negatives (a lesion free area that received visual dwell but was not reported and therefore was assumed to be normal).

In order to determine the spectral frequency decomposition of each region, they were processed using a filter bank that contained quadrature-mirror filters, in a process known as Wavelet Packets. During the decomposition of each region the signal energy in each spatial frequency band was calculated. Each band was represented by a combination of two numbers, one that indicated where in the tree the band was located (levelwise) and one that indicated if the signal being processed had been low-(or high-) passed in the previous step and was now being low-(or high-)passed.

Once the energy profiles for each region were determined, the regions were split in two sets. One set was used to train an Artificial Neural Network regarding each radiologist's decision patterns; the other set was used to test the network, namely, to predict which type of decision outcome would such profile generate, for each observer. The ANN chosen was an Adaptive Ressonance Theory network with 21 features as input – 20 mean energy values, per spatial frequency band, and one number, from 1 to 4, which indicated which observer was reading that profile. The system is shown in Figure 1.

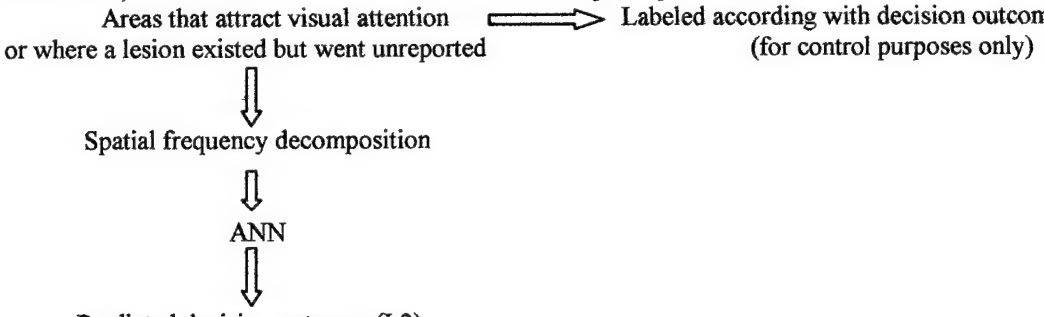


Figure 1. Basic flowchart of Computer Assisted Perception

Thus, to calculate the error generated by CAP, L1 is compared to L2. If L1 = L2, then the error is zero. If not, the error counter is incremented by one unit. The final error, per decision outcome, equals the error counter divided by the total number of samples of each decision outcome.

3.Results

Taking into account the individual signatures generated by each decision outcome made by the observers (N=17,080), it follows that:

- FNs yield lower enegy profiles than TPs in 10 spatial frequency bands;
- FPs yield lower energy profiles than TPs in 4 spatial frequency bands;
- TNs yield higher energy profiles than TPs in 11 spatial frequency bands;
- TNs also yield a higher energy profile than FNs in 4 spatial frequency bands;
- TNs yield lower energy profiles than FPs in 2 spatial frequency bands;
- FNs yield lower energy profiles than FPs in 6 spatial frequency bands.

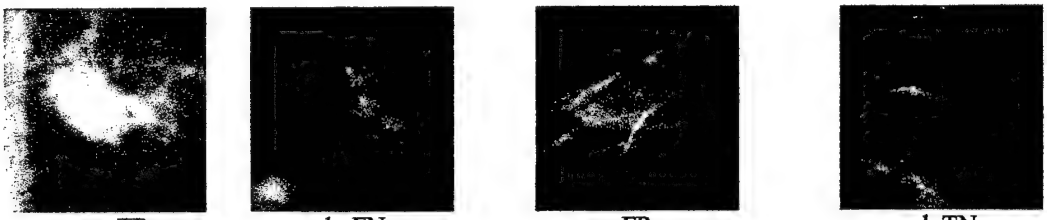
Statistical significance was tested by analysis of variance and Scheffe's test ($p<0.05$). All of the aforementioned results were statistically significant.

Additionally, it is possible to separate the differences according to lesion type. In this way,

- For masses: 6 spatial frequency bands contribute to significantly differentiate TPs from FPs; 4 bands contribute to differentiate FNs and FPs; 11 bands contributed to differentiate TNs and TPs, 2 bandsto differentiate TNs and FPs and 4 bands to differentiate TNs and FNs.
- For calcifications: 12 spatial frequency bands contribute to significantly differentiate FNs from TPs; 11 to differentiate TNs and TPs; 2 bands to differentiate TNs and FPs and 4 bands to differentiate TNs and FNs.

Figure 2 shows a sample of each type of decision outcome, and the spatial frequency profile that they generate. In this figure, the lesion designated by TP was indicated by all observers; the one labeled FN was missed by all observers; the

one labeled FP was incorrectly indicated by 3 observers, and the area designated as TN did receive visual dwell, but correctly was not indicated as containing any lesion.



a. TP b. FN c. FP d. TN

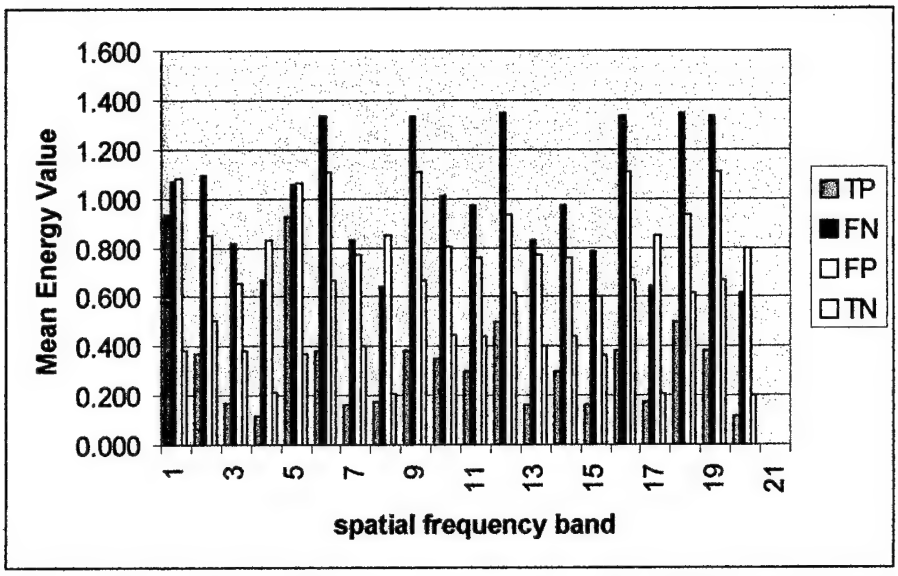


Figure 2. Typical regions used for processing and the energy profile that they yield.

In Figure 2 it is shown that some bands have a high mean energy value, whereas others have a small value. If all the bands are normalized by the average energy (per spatial frequency band) of the regions selected for processing, regardless of decision outcome, the mean results for all regions used in this experiment are shown in Figure 3.

Using half of the energy profiles to train CAP and half to test it, yield the following rates of correct predictions:

- FPs: 81%
- FNs: 64%
- TNs: 71%
- TPs: 55%

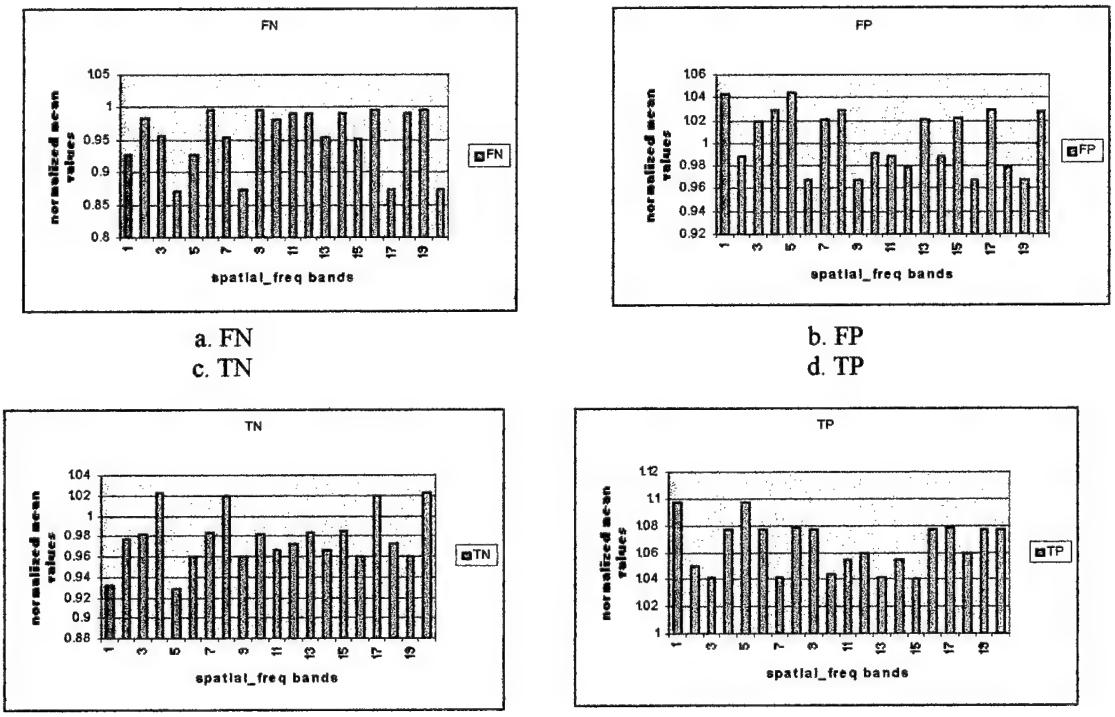
4.Discussion

The results shown herein indicate that different decision outcomes possess a characteristic signature, in the spatial frequency domain, and that the information contained in this signature alone is sufficient to predict which type of decision outcome a given observer will make in that area of the image. Furthermore, pairs of decision outcomes can be discriminated based upon the information contained in a finite, and often small, number of spatial frequency bands. This has significant consequences for the distinction of microcalcifications that lead to detection (thus yielding a True Positive decision) versus the ones that are not detected (yielding a False Negative decision), because these two types of microcalcifications are different in 12 spatial frequency bands. In this way, it is conceivable that CAP could potentially be used to flag to the observer regions that obbey the profile of FN calcifications, thus improving the observer's performance. This could be done by comparing in real time the labels generated by CAP and the decisions made by the observer when examining a given mammogram.

Additionally, for masses, there are differences in 6 spatial frequency bands between areas that yield a correct detection of a mass (a TP decision) versus the ones that yield an incorrect decision (a False Positive decision). These, in turn,

are different from the other type of incorrect decision, the False Negatives, in 4 spatial frequency bands. Again this goes to show CAP's potential in helping improve observer's performance by feeding back to the observer the likelihood that a given decision is correct or not.

Interestingly, there were no statistically significant differences between the masses that were not reported (FNs) and the ones that were correctly reported (TPs), indicating that to the proposed system the spatial signature of both types were similar. In this case it is hard to say if CAP considered all masses to be very subtle (FNs) or if it considered them to be well defined (TPs).



a. FN
b. FP
c. TN
d. TP

Figure 3. Mean energy profiles that characterize each decision outcome.

The error rates generated by CAP are consistent with the view that part of the noise is in the image and part is in the observer. Thus, by just taking into account the image noise and a fixed measure of the observer's noise (given by the number from 1 to 4 which identified the observer) CAP was able to correctly predict 81% of the False Positives and 64% of the False Negatives made by the observers. These percentages of error prediction could have a significant impact in the clinical practice, if CAP was used to provide feedback to the observers in real-time. We are currently investigating this hypothesis.

5. Acknowledgement

This work was partially supported by Grant DAMD17-97-1-7103 between the USAMRMC and C. F. Nodine.

6. Reference

1. M. Giger and H. MacMahon. Computer-aided diagnosis. Radiologic Clinics of North America, 34:565-596, 1996.

2. E. A. Krupinski and C. F. Nodine. Gaze duration predicts the location of missed lesions in mammography. In A. G. Gale et al., editor, Digital Mammography, Elsevier Science B.V., 1994.

3. J. G. Elmore, M. B. Barton, V. M. Moceri, S. Polk, P. J. Arena and S. W. Fletcher . Ten-year risk of false positive screening mammograms and clinical breast examinations. New England Journal of Medicine 338(16):1089-1096, 1998.

PROGRESS IN BIOMEDICAL OPTICS AND IMAGING

Vol. 1, No. 26
ISSN 1605-7422

Reprinted from

Medical Imaging 2000

Image Perception and Performance

16–17 February 2000
San Diego, California

Proceedings of SPIE
Volume 3981

Do Subtle Breast Cancers Attract Visual Attention During Initial Impression?

Calvin F. Nodine, Claudia Mello-Thoms, Susan P. Weinstein, Harold L. Kundel, and Lawrence C. Toto
University of Pennsylvania, Philadelphia, PA 19104-6086

ABSTRACT

Women who undergo regular mammographic screening afford mammographers a unique opportunity to compare current mammograms with prior exams. This comparison greatly assists mammographers in detecting early breast cancer. A question that commonly arises when a cancer is detected under regular periodic screening conditions is whether the cancer is new, or was it missed on the prior exam? This is a difficult question to answer by retrospective analysis, because knowledge of the status of the current exam biases the interpretation of the prior exam. To eliminate this bias and provide some degree of objectivity in studying this question, we looked at whether experienced mammographers who had no prior knowledge of a set of test cases fixated on potential cancer-containing regions on mammograms from cases penultimate to cancer detection. The results show that experienced mammographers cannot recognize most malignant cancers selected by retrospective analysis.

Keywords: Visual attention, Missed cancers, Retrospective analysis, Eye fixations

1. INTRODUCTION

Should detected breast cancers that can be seen retrospectively on the immediately prior mammogram be considered missed or incident cancers?

This is a difficult question to answer because perceptual knowledge of lesion features and location bias the observer's interpretation in retrospectively looking for the cancer on the prior mammogram. The issue of missed cancers is a major source of malpractice lawsuits filed against radiologists (Berlin, 1999). This is in spite of the fact that even an "expert" making a retrospective analysis cannot neutralize apriori knowledge in viewing the radiographic image after having once recognized the cancer (Berlin, 1996).

Our experiment looked at subtle cancer cases. These consisted of: (a) a group of 20 subtle cancer cases that were not reported on the mammogram immediately prior to detection (mean interval = 14.25 mo.), but were visible in retrospect when analyzed by an experienced breast imager (SPW); (b) a group of 10 true incident cancer cases, and, (c) 10 cancer-free cases (2-year follow up). These subtle cancer cases were digitized to 50 micron image resolution. The image gray scale for each view was automatically set by a linear Look-Up-Table (LUT) algorithm in which a binary version of the original image was used to find the breast outline, and then the intensity range within the original breast image segment was sampled to define the LUT .

In order to design a fair test of the question, we needed to choose observers who were experienced mammographers, but who did not have apriori knowledge of the mammogram cases in the test set. They were, however, given information indicating that they would be seeing subtle lesions, so their suspicion was raised. In addition, we monitored eye position during initial interpretation of the mammograms in order to provide an objective measure of whether or not the subtle lesions were looked at (fixated) independently of being reported. When the initial interpretation was concluded, the observer gave a general impression (normal or abnormal). Without interruption, the observer was given additional viewing time to examine the mammogram case using full-resolution digital zoom. If a potentially malignant or suspicious lesion was recognized, the observer localized it with a mouse cursor. This action called up a menu prompting the observer to classify the lesion by type and give a decision-confidence rating. If no suspicious findings were found, the observer terminated the trial by calling up the next image. This resulted in a default normal decision.

The focus of my paper is on how analysis of eye-position data are related to whether or not subtle lesions are fixated long enough for the observer to make a decision about them, and how these data are related to initial decision outcome and subsequent zooming analysis. The complementary paper to this (3981-25) presented by Claudia Mello-Thoms will focus on how zooming data are related to localizing subtle lesions and how these data are related to eye position and final diagnostic decision outcomes. It should be noted that these two papers are based on the same experiment.

In *Medical Imaging 2000: Image Perception and Performance*, Elizabeth A. Krupinski, Editor, Proceedings of SPIE Vol. 3981 (2000) • 1605-7422/00/$15.00

2. MATERIALS AND METHODS

We recorded eye-position data (ASL, Model 4000SU, Bedford, MA) on 4 experienced breast imagers viewing a test set consisting of 20 retrospectively visible cancer cases not reported on initial screening (NR), 10 prospectively reported cancer cases (R), and 10 cancer-free cases. Two mammographic views, CC and MLO, were digitized for each case and displayed on a 21" high-resolution (2560 x 2048) workstation (Orwin, Model DS5000L, Amityville, NY). This was no ordinary workstation in that a data record was generated on each observer which contained: event times of mouse clicks indicating decision events; lesion locations; eye-fixation locations; zoom locations; and, zoom durations for each mammographic view of each case.

3. RESULTS

Did experienced breast imagers look at subtle lesions long enough to recognize malignancy? We used 1000 ms as the dwell threshold for recognizing a breast lesion based on earlier work in which we showed that a minimum of 1000 ms was required for a positive decision (Krupinski, Nodine, Kundel, 1998).

Considering that NR lesions are true cancers, the answer to the question that prompted this study is "yes". Initially, 66 % of NR lesions v.60 % of R lesions were fixated for >1000 ms.

Phase 1 time was highly correlated with total number of fixation clusters as shown in Figure 1 which relates total number of cumulative clusters per image to phase 1 viewing time. This suggests that most visual search time was spent focally searching and examining image features for possible lesions. This is the effect of "zooming" with the eye.

Insert Fig. 1 here

How does fixating relate to initial decision outcome? Initially, observers over reported as positive 69% of NR and 81% of R test cases. Bar Graph 1 shows the yield of decision outcomes resulting from the initial decision for fixations >1000 ms. for NR and R test cases. Only slightly more than half of NR cases (58%) were correctly interpreted compared 76% of R cases based on initial decision.

Insert Bar Graph 1 here

False positive rates of 28% and 19% are not too far out of line given that in clinical practice, for patients recommended for biopsy, only 1 in 3 will typically have a cancer. But we did not allow observers to perform additional imaging evaluations in the present study.

How long did observers fixate to generate a decision? Observers were suspicious since they were told that they were looking for subtle lesions. Initially, experts eyes fixated subtle lesions, but they had difficulty recognizing true from false malignant lesions. In reality mammographers do not rely on 2 mammographic views alone to determine malignancy, but follow up with additional evaluation images such as mag views, ultrasound and ultimately biopsy.

Mean fixation cluster dwell times by decision outcome for NR and R test cases are shown in Bar Graph 2.

Insert Bar Graph 2 here

Interestingly, decisions with mean dwell times >1000 ms. (n=166) were 7 times longer than the corresponding decisions with mean dwell times <1000 ms. (n= 204).These latter decision times ranged from 372-680 ms. suggesting that 1000 ms. is a good dwell threshold for defining "directed attention".

Bar Graph 2 is based on a lesion analysis of CC and MLO views using truth table generated by the breast imager (SPW). The long dwells, especially for NR test cases, suggest difficulty differentiating true from false malignant lesions, implying a low signal-to-noise ratio. These average dwell times are consistent with previous studies (e.g. Krupinski, Nodine, Kundel, 1998)

Does fixating a potential lesion result in subsequent zooming of it? Yes, 80% of initially fixated lesions were subsequently zoomed. No difference between NR and R.
Fixations that were subsequently zoomed resulted in longer dwells (1884 ms) than fixations that were not subsequently zoomed (1224 ms, $p<.05$, Scheffe test) indicating that findings that captured visual attention were followed up by zooming.

4. DISCUSSION

Experienced breast imagers with high suspicion initially failed to recognize 42% of retrospectively visible subtle malignant breast lesions. Does this mean that these subtle lesions should not be considered "missed cancers" but rather true incident cancers because they could not be differentiated from normal background structures? Probably not.

We have acknowledged the high rate of false positives in this study and attributed it, in part, to increased suspicion on the part of the observers. It is also due to the scoring of overall performance which was done on a lesion basis meaning that observers could, and did, generate FPs on both CC and MLO views. They also got credit for finding cancers on both views. But, from a clinical standpoint, the troubling aspect of this performance was not the high false positive rate, but the higher miss rate.

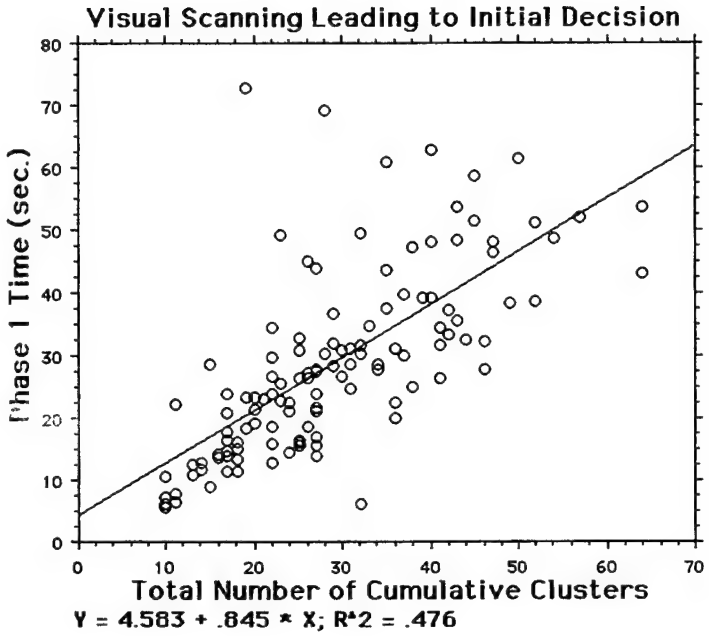
Why most cancers in the NR cases were not recognized at the initial viewing is unclear, but three thoughts come to mind. First, mammographic images are far from perfect and this study used digitized versions which may have degraded the signal-to-noise ratio. However, although subtle, the lesions were retrospectively visible on the digitized version. Second, we did not give the mammographers the option to further evaluate the areas of potential cancer. They knew that they had to rely on the 2 views supplemented only by full-resolution zooming for a malignant/non-malignant interpretation. This is not the way experienced mammographers work in practice and may have played out by a higher than normal miss rate. Finally, I hope this study puts a nail in the "retrospective analysis" coffin. Although a lesion may be visible in retrospect, our experienced breast imagers had extreme difficulty differentiating true positive lesions from false positive ones, even though they were alerted to the presence of "subtle" cancers. So much for expert testimony based on retrospective analysis. It is easy to detect a subtle cancer with the benefit of apriori knowledge, but without it even highly experienced breast imagers stumble.

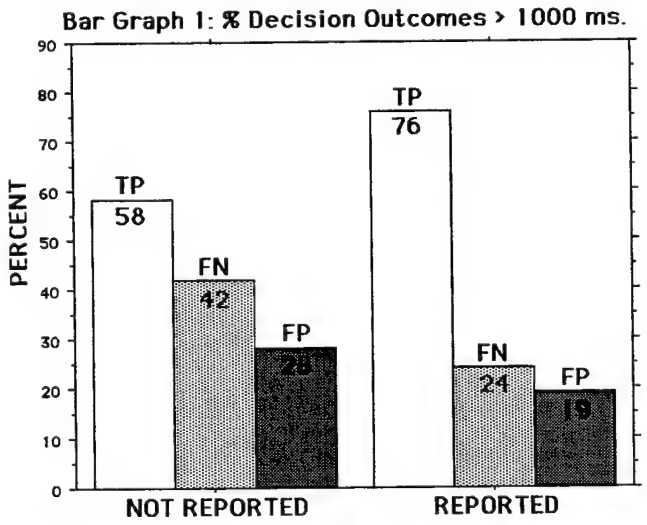
5. ACKNOWLEDGEMENTS

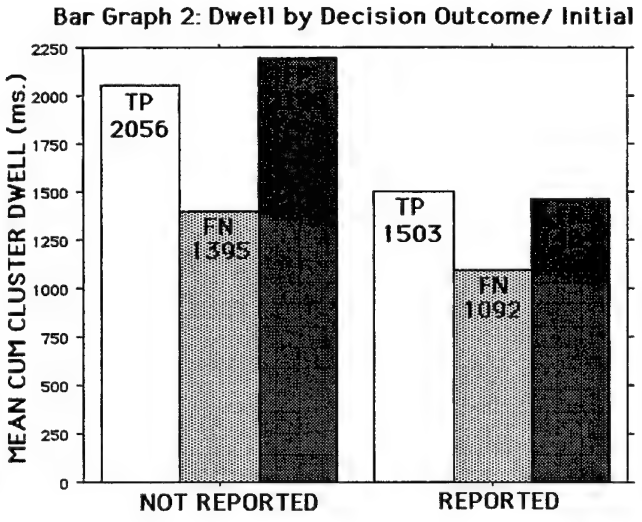
This research was supported in part by DAMD17-97-1-7130

6. REFERENCES

1. Berlin L. The missed breast cancer: Perceptions and realities. ARJ 1999;173:11-61-1167.
2. Berlin L. Perceptual errors. ARJ 1996;167:587-590.
3. Krupinski EA, Nodine CF, Kundel HL. Enhancing recognition of lesions in radiographic images using perceptual feedback. Opt Eng 1998;37: 813-818.


Visual Scanning Leading to Initial Decision
Phase 1 Time (sec.)
0
10
20
30
40
50
60
70
80
Total Number of Cumulative Clusters
Y = 4.583 + .845 * X; R^2 = .476


Bar Graph 1: % Decision Outcomes > 1000 ms.
PERCENT
0
10
20
30
40
50
60
70
80
90
TP
58
FN
42
FP
28
TP
76
FN
24
FP
19
NOT REPORTED
REPORTED


Bar Graph 2: Dwell by Decision Outcome/ Initial
MEAN CUM CLUSTER DWELL (ms.)
0
250
500
750
1000
1250
1500
1750
2000
2250
TP
2056
FN
1395
TP
1503
FN
1092
NOT REPORTED
REPORTED

PROGRESS IN BIOMEDICAL OPTICS AND IMAGING

Vol. 1, No. 26
ISSN 1605-7422

Reprinted from

Medical Imaging 2000

Image Perception and Performance

16–17 February 2000
San Diego, California

Proceedings of SPIE
Volume 3981

Image Structure and Perceptual Errors in Mammogram Reading: A Pilot Study

Claudia Mello-Thoms①,②, Stanley Dunn②, Calvin F. Nodine① and Harold L. Kundel①
①University of Pennsylvania School of Medicine, Department of Radiology
②Rutgers University, Department of Biomedical Engineering

Abstract

Early detection of breast cancer is very desirable, considering that it can significantly change the prognosis for a woman diagnosed with this disease. Nonetheless 10-30% of all breast cancers are missed by the radiologist, albeit they are visible in the mammogram. In this work we have studied the underlying structure of the image in the location of the lesions that were missed and the ones that were found, as well as in the locations of the lesions that did not exist but were reported by the radiologist. We have shown that there is a statistically significant difference in the information content of different frequency bands that results in various decision types. We have also shown that it is possible to use a pattern classifier, based upon the information contents of the spectral decomposition of a local image region, to predict the most likely decision outcome.

Keywords: Image structure, perceptual errors, mammogram reading, wavelet packets.

1.Introduction

Early detection can significantly change the prognosis for a woman diagnosed with breast cancer. Thus, renewed efforts have been made to develop accurate imaging techniques that can detect abnormalities of smaller sizes. Nonetheless, a problem that is usually overlooked when considering such imaging techniques is the radiologist's ability to correctly interpret what is on the image. It has been shown [1] that 10-30% of all breast cancers are missed, being only found retrospectively, albeit they are visible in the mammogram. Furthermore, from these, 65% are fixated by the high-resolution central/foveal vision [2]. In other words, these cancers are not missed because of search errors, but because of perception and decision-making errors.

Kundel and Nodine [3] have derived a model that links perception and decision making in medical image reading. This model predicts that perception, and ultimately decision making, start out with a global impression of what is in the image. This global impression is compared to a cognitive schema, stored in memory, of similar images that the observer has seen in the past. This comparison flags regions of potential abnormality, which the observer examines by visually scrutinizing the area with the high-resolution fovea. This results in the extraction of features that are processed and used for object categorization. If a positive fit is found with some representation in memory, additional visual search is performed, until an internal threshold is crossed, and the abnormality is decided positive or negative.

Many factors have been shown to play a role in aiding or preventing lesion detection. Among these, the relationship between the abnormality and the background tissue surrounding it has been shown to be one of the most important. Burgess and colleagues [4] have shown that, in mammograms, lesion detectability is not related to the size of the lesion, but rather to a power law which takes into account the signal energy and the background structure power spectrum. This means that even large lesions can be missed, if certain conditions hold between the lesion and its surrounding tissue.

In this paper we will examine the relationship between breast masses and their surrounding tissue as a function of what decision type they yield, namely, if they yield True Positives (that is, the observer correctly finds a malignant abnormality present in the mammogram), True Negatives (if the observer correctly interprets normal tissue as being lesion-free), False Positives (when the observer incorrectly interprets normal tissue as being malignant) and False Negatives (when the observer fails to indicate a malignant lesion that is visible in the mammogram).

2.Materials and Methods

Eight experienced observers (3 mammographers from the staff of the Hospital of the University of Pennsylvania, HUP, and 5 fellows undergoing training at the same institution) read 5 two-view (cranio-cauldal, CC, and medio-lateral-oblique, MLO) mammogram cases. All cases had a malignant mass visible in at least one view. One case contained multiple malignant masses, visible in both views. These cases were obtained from the archives of HUP. The films were digitized using a Lumiscan Model 100 digitizer (Lumisys Inc, Sunnyvale, CA), using a 100 microns spot size. The two-views were displayed

In *Medical Imaging 2000: Image Perception and Performance*, Elizabeth A. Krupinski, Editor, Proceedings of SPIE Vol. 3981 (2000) • 1605-7422/00/$15.00

on a single 19-inch, 2048x2048 gray scale monitor (GMA 201, Tektronix, Beaverton, OR), interfaced to a Sun Sparc computer (Sun Microsystems, Sunnyvale, CA).

The observers were instructed to search for malignancy, and freely examined the cases until they felt confident to point out if and where a malignant lesion was present. The eye position of the observers was monitored during search, and it was used to determine the areas in the image that attracted the observers' attention.

The eye position of each of the observers was played back over the mammogram case examined, and from each case 10 regions were manually extracted using a mouse-controlled cursor. These regions contained true lesions that were indicated by the observer (and were labeled TP), true lesions that were missed by the observer (labeled FN), lesion-free areas that were indicated by the observer as being lesion-containing areas (FP) and lesion-free areas that were correctly interpreted by the observers as being normal tissue (TN).

Each of these regions was processed using a filter bank that contained quadrature-mirror filters, using a process known as Wavelet Packets. This tree was two-levels deep. During the decomposition of each region some statistical parameters were calculated for each frequency band, including the mean and standard deviation and the signal energy in that band. Each band was represented by a combination of two numbers, one that indicated where in the tree the band was located (levelwise) and one that indicated if the signal being processed had been low-(or high-)passed in the previous step and was now being low-(or high-)passed.

3. Results

The mean values for the energy in the different frequency bands is listed in Table 1.

band	Mean energy value		band	Mean energy value
00	11327.45		22	2.08
01	4.29		23	1.99
02	20.17		30	16.11
03	25.01		31	2.08
10	42502.70		32	59.05
11	0.44		33	0.44
12	16.11		40	19.86
13	19.86		41	1.99
20	0.44		42	0.44
21	11.38		43	73.77

Table 1. Mean values for the energy per frequency band.

As shown, there is a wide variability in the information contents of each of the bands. Thus, the bands were divided in three classes: the low energy (which had a mean <= 10), the medium energy (10 < mean <= 50) and the high energy bands (mean > 50). Thus, in order to assess the contribution of each band on the decision outcomes an ANOVA analysis was run. In all Scheffe tests listed below, the significance level was 5%. Table 2 lists these results.

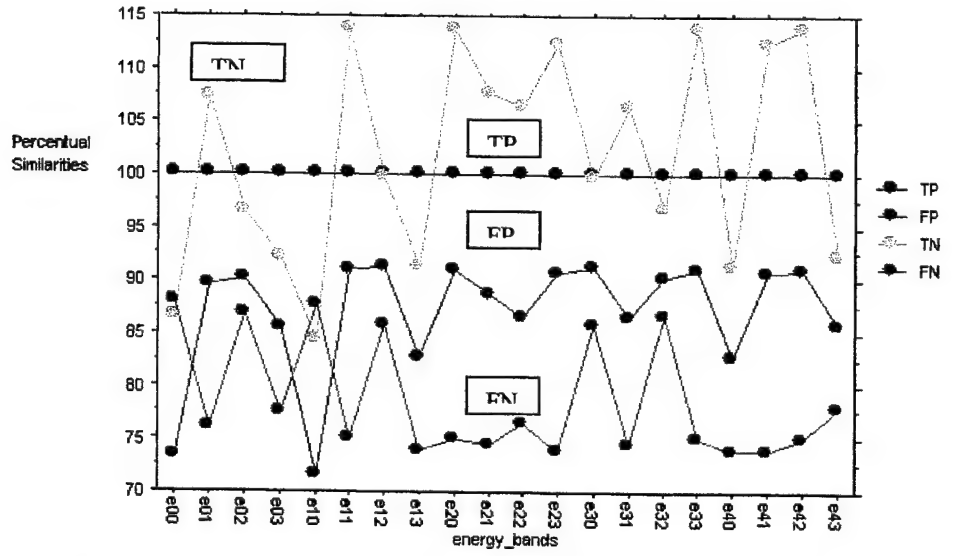
Band class	Band Name	Contributed for the differentiation between:
High	00 10	TPs from FPs ($p < 0.05$) TPs from FPs ($p < 0.05$)
Medium	21	FPs from TNs ($p < 0.05$)
Low	42	FPs from TNs ($p < 0.05$)

Table 2. List of the energy bands that contributed to the differentiation of pairs of decision outcomes, as tested using Scheffe's test.

The importance of these results stems from the fact that they clearly state that there exist differences in the energy contents, per frequency band, of the regions of the image that result in different decision outcomes.

Furthermore, if one considers the mean values of energy on the different frequency bands that lead to True Positive decision outcomes as a base value, then the breakdown of energy, in percentual values, relative to the levels of the TPs, for the remaining decision outcomes is shown in Figure 1.

Figure 1. Percent differences between the TPs and FPs, FNs and TNs.



As Figure 1 shows, the energy contents are generally higher for the True Negatives, particularly for the intermediate energy bands. As their contents begin to change, that is, to lose power in the majority of the bands, the False Positives are formed. As power continues to decrease, the False Negatives come about. An ANOVA was run on these percentual differences, and it was found that there are statistically significant differences between FNs and FPs (Scheffe test, $p < 0.05$), FNs and TNs ($p < 0.05$) and FPs and TNs ($p < 0.05$).

Knowing from the first ANOVA which bands are responsible for the differentiation between different pairs of decision outcomes, we decided to use a Neural Network to predict, from the values of the energy in the high- and intermediate-bands, the decision outcome that that a particular region of image would yield. Additionally a parameter that ranged from 1 to 8 was used to inform the network which observer had provided the data being examined. The reasoning here is that different observers may perceive the same region of the image very differently; for example, a more experienced observer may be able to detect a subtle mass whereas another observer may not see anything.

Using an Adaptive Resonance network the results shown on Table 3 were obtained, in terms of correct and incorrect responses. Once more, the purpose of the network was not to determine if an abnormality was present or not on a particular region of the image, but rather, to determine which decision outcome was more likely for a given observer when examining that region of the image.

Class	Correct Predictions	Incorrect Predictions
TP	44/69 = 64%	25/69 = 36%
FP	46/60 = 77%	14/60 = 23%
TN	37/53 = 70%	16/53 = 30%
FN	5/18 = 28%	13/18 = 72%

Table 3. Percent values for correct and incorrect decision outcomes as predicted by the neural network.

This result clearly indicates that it is possible to separate TPs, TNs and FPs based upon the energy decomposition of the region indicated by the observer. Nonetheless, the results for the False Negatives were not good. This is certainly a reflection of the limited number of such samples that was available in this pilot study.

4.Discussion

These results indicate that there is a particular configuration of energy, in the frequency domain, that leads observers to detect true lesions. Furthermore, there also exist particular energy configurations that will likely lead the observers to make False Positives, False Negatives and True Negatives.

When using a pattern classifier to automatically predict which decision outcome will a particular combination of energy in different frequency bands yield, we found that the TPs, FPs and TNs could be reliably predicted, but, due to the small sample size, the same was not true for the FNs. We believe that as our research proceeds, with a much larger database, the results for the FNs will significantly improve.

5.Acknowledgement

This work was partially supported by Grant DAMD17-97-1-7103 between the USAMRMC and C. F. Nodine.

6.Reference

1. M. Giger and H. MacMahon. Computer-aided diagnosis. Radiologic Clinics of North America, 34:565-596, 1996.

2. E. A. Krupinski and C. F. Nodine. Gaze duration predicts the location of missed lesions in mammography. In A. G. Gale et al., editor, Digital Mammography, Elsevier Science B.V., 1994.

3. H. L. Kundel. Perception and representation of medical images. In Proceedings of the SPIE, Image Processing, vol 1898, pp 2-12, 1993.

4. A. E. Burgess, F. L. Jacobson and P. F. Judy. On the detection of lesions in mammographic structures. In E. Krupinski, editor, Proceedings of the SPIE Conference on Image Perception and Performance, vol 3663, pp 304-315, 1999.

PROGRESS IN BIOMEDICAL OPTICS AND IMAGING

Vol. 1, No. 26
ISSN 1605-7422

Reprinted from

Medical Imaging 2000

Image Perception and Performance

16–17 February 2000
San Diego, California

Proceedings of SPIE
Volume 3981

An Unobtrusive Method for Monitoring Visual Attention During Mammogram Reading

Claudia Mello-Thoms, Calvin F. Nodine, Susan P. Weinstein, Harold L. Kundel and Lawrence C. Toto
University of Pennsylvania School of Medicine, Philadelphia, PA 19104

Abstract

The use of feedback to the observer of the regions of the image that attract prolonged visual dwell (> 1000 ms) has been shown to improve nodule detection performance in reading chest x-rays. The application of such a feed back mechanism in mammography seems appropriate, but it is often discouraged by the inherent difficulties of using an invasive eye-tracking system. In this paper we discuss the use of an alternative method, namely, a digital zoom window, to monitor where the observer's attention is focused on the image. We have shown that the order in which the zooms occur, as well as the duration of certain zooms, is statistically correlated with decision outcome for a given region of the image. Furthermore we show a strong correlation between zooming and prolonged fixation.

Keywords: Breast cancer, eye-position monitoring, zoom window, decision outcome.

1. Introduction

Mammography is the standard screening test for breast cancer. Nonetheless, sensitivity of Mammography is about 85-90%. A question that naturally arises is: these cancers were missed due to faulty search or recognition failure?

Eye position studies have shown that the majority of missed cancers are in fact looked at [1], and the dwell times on these locations are almost as long as on the cancers that are reported. Furthermore, eye position and a dwell time threshold have been used to provide feedback to observers about the locations of possibly missed nodules in chest x-ray readings, and detection performance has improved as a result [2]. Thus, in order to improve breast cancer detection, one interesting alternative is the application of perceptual feedback. Unfortunately, eye position monitoring, using an eye-tracker, is a cumbersome and intrusive research tool. It suffers from a variety of drawbacks, such as the need to keep the calibration updated, adjust for spurious reflections from the observer's eye glasses or contact lenses and reflections from the observer's skin, difficulty in tracking the pupil if the observer tends to lower his or her eyelids, etc. Furthermore, even with a perfect observer there is still some discomfort due to eye dryness and headaches caused by the infrared beam used to monitor limbus reflections. Thus, it is impractical to use such a system for long-term monitoring of the observers' attention when reading medical images.

In their daily practices mammographers read mammograms using a two-pass strategy. In the first pass they globally search the mammogram for typical abnormalities, and in the second, using a magnifying lens, they repeat the search, looking for microcalcifications or other subtle findings. In this way, we hypothesized that by allowing them to use a digital zoom, when reading mammogram cases on a computer workstation, we would be able to monitor where on the image their attention was directed without using any eye position monitoring. Furthermore, we hypothesized that the length of the time that the zoom window is stationary at a particular location is related to the decision outcome yielded at that location, just as the visual dwell is related to decision outcome.

In this paper we will examine the use of this digital zoom window to monitor the experts' visual attention, and compare the results with an eye-tracking system. We will show how the use of the zoom is related to the decision outcomes in a task where the observers are instructed to search for malignancy. We will also show which percentage of these responses had initially attracted the attention of the observers, during a scanning phase in which eye position was monitored using an eye-tracking system, and how many of them were further investigated on a second phase in which the observer was allowed to zoom in onto a region of interest in the image.

2. Materials and Methods

Four experienced observers (2 staff mammographers and 2 fellows undergoing training at the Hospital of the University of Pennsylvania) examined 40 two-view mammogram cases on a digital workstation. These 40 cases were obtained from the archives of the same hospital by one of the authors (SPW), who is a mammographer but did not participate in the study. These cases contained 10 cancer-free cases which had been stable for a period of 2 years (N); 10 cases in which the malignant lesion present had been reported (R), and 20 cases in which a malignant lesion, albeit present and visible, had not been reported, being only found retrospectively (NR). Malignancy for all lesions was determined through biopsy.

The films were digitized using a Lumiscan Model 100 digitizer (Lumisys Inc, Sunnyvale, CA), with a 50 microns spot size. The two-views were displayed on a single 21-inch, 2560x2048 gray scale monitor (ORWIN Associates, Amityville, NY), interfaced to a Gateway GP6-266 computer (Gateway, North Sioux City, SD) running Windows 95 (Microsoft

In *Medical Imaging 2000: Image Perception and Performance*, Elizabeth A. Krupinski, Editor, Proceedings of SPIE Vol. 3981 (2000) • 1605-7422/00/$15.00

Corporation, USA). The cranio-caudal view was displayed on the left-hand side of the display, and the medio-lateral oblique view was displayed on the right-hand side.

The observers were instructed to search for malignancy. The experiment was divided into two phases. In the first phase the observers visually searched the images until they felt confident to provide an initial impression about the case, namely, if it was normal or abnormal. During this phase the eye position was tracked using an infrared based system, the 4000SU (Applied Science Laboratories, Bedford, MA). This system has an accuracy of about 1°. Once the observer concluded if it was a normal or abnormal case, they were instructed to pull down a menu on the screen, where they gave their initial impression about the case. This marked the end of the first phase, and the eye-tracking system was turned off for the second phase, in which the observers freely used a digital zoom window to further study any areas of the image where they suspected that a malignant lesion was visible. This zoom window was about 401 x 401 pixels wide, and it was centered at the location indicated by the observer using a mouse-controlled cursor. Inside the zoom window the image was seen at its original resolution of 50 microns. In this phase they were instructed to, upon detecting a malignant lesion, place a mouse-controlled cursor over the center of the lesion and click. This action would prompt a menu to appear in the screen, where they answered what type of abnormality they had found (mass, calcification, architectural distortion) and how confident they were that it was indeed malignant (low, medium and high confidence). These responses were saved to a file that also contained information regarding the time when the decision was made.

Unbeknownst to the observers, the locations and the duration of the zooms, as well as their sequence, were also recorded to a file. This allowed us to keep track of the areas that attracted the observers' attention, and also how conspicuous a stimulus element had to be in order to be zoomed (that is, were the most conspicuous elements zoomed in first?).

Based upon knowledge provided from pathology reports and posterior films, where the cancer was reported, one of the authors (SPW) marked the coordinates of all of the lesions present in this test set and determined their nature (mass, calcification, architectural distortion). This data allowed us to build a truth table, against which we compared the observers' assessment, and rated their decision outcomes as being True Positives (TPs), False Positives (FPs), True Negatives (TNs) and False Negatives (FNs).

3. Results

In this section we will present the results of using the digital zoom window, and compare these with data generated by the eye-tracking system in the first phase of the experiment. Eye fixations were clustered by grouping raw points of eye position data using certain rules. For example, to be included in a cluster the points had to occur in sequence and in the same neighborhood. Furthermore, the fixations had to fall within a grouping that did not exceed 2.5 degrees. If the distance was greater than that, a new cluster was created, having as center the centroid of that group of fixations. For each cluster, the dwell time on the location of the cluster was calculated by multiplying the total number of data points inside that cluster by 1/60, which is the sampling rate for the ASL system.

3.1. Comparing the use of the zoom window with the clusters

In order to compare the two measures of observers' attention, namely, the use of the zoom window and the clustering of eye position, we calculated the mean number of zooms and clusters per case type. Furthermore, we also measured the percentage of clusters (> 1000ms) that were later zoomed, as well as the percentage of zooms that occurred in locations where a cluster (> 1000ms) existed during the scanning phase. This is shown on Table 1.

Case Type	Mean # of Clusters	Mean # of Zooms	% of Clusters that were later zoomed	% of Zooms that occurred in locations of clusters
R	10.593	6.767	30	76
N	7.680	3.000	26	86
NR	9.748	5.673	29	73

Table 1. Average use of the zoom window in comparison with the visual dwell clusters

3.2. Comparing zoom and dwell times during search with the responses made

Table 2 relates the decisions made by the observers, per image type, with the locations that attracted their visual attention during phase 1 and the ones in which they zoomed on phase 2. As it is clear from this table, most of the locations that elicited a response from the observers were either zoomed or received significant (>= 1000 ms) visual dwell during the scanning phase. Furthermore, the differences in the percentages of the decisions that attracted visual attention during the phase 1 from the ones that were zoomed on phase 2 was not statistically significant.

Image Type	Decision Outcome	Location received long (>1000ms) visual dwell	Location was zoomed on Phase 2
R	FN	29%	58%
	FP	42%	77%
	TP	71%	89%
NR	FN	46%	56%
	FP	67%	91%
	TP	54%	91%
N	FP	25%	88%

Table 2. Relationship between the areas in the images that yield a decision outcome and the percentage of them that were either looked at, during the first phase, or zoomed in, during the second phase of the experiment.

3.3. Effect of zoom order on decision type

In order to assess if the most conspicuous elements were zoomed in early or late during the zooming phase, we have numbered the zooms according with the order in which they occurred, and we have related this order to decision outcome. This is shown in Table 3.

Case Type	Zoom Number	Decision Outcome	Statistically Significant Difference Yielded
R	1st	TP	Between FN and TP (Scheffe's test, $p < 0.05$)
	2nd	TP	
	3rd	FP	Between FP and TP ($p < 0.05$)
	4th	FN	
N	1st	TN	There were no statistically significant differences
	2nd	FP	
NR	1st	TP	Between FN and TP ($p < 0.05$)
	2nd	FP	Between FP and TP ($p < 0.05$)
	3rd	FN	

Table 3. Relationship between order in which the zoom occurred and the decision outcome that it yielded, as well as the statistically significant differences between the decision outcomes., as measured by zoom order.

3.4. Effect of zoom length, per zoom number, on decision type

Considering that the order in which the zooms occurred was directly related to the decision outcomes, we decided to verify if the duration of the zoom, measured by how long the observer kept the zoom window fixed in one location, had any significant correlation with the decision outcomes.

For the R cases, for the eighth zoom, there were statistically significant differences between FNs and FPs ($p < 0.05$) and between FPs and TPs ($p < 0.05$).

For the N cases, for the first zoom, it lasted about 4 seconds when it yielded a TN decision, whereas it lasted about 9 seconds when it yielded a FP. This difference was statistically significant ($p < 0.05$).

For the NR cases, for the second zoom, there were statistically significant differences between FNs and FPs. In this case the FNs lasted about 3 seconds, whereas the FPs lasted about 9 seconds.

When comparing these numbers with those yielded by the visual dwell, during the phase 1, on the locations where later the observers indicated (or fail to do so) the presence of a malignant lesion, there were no statistically significant differences for the R and N cases. For the NR cases, there was a statistically significant difference was between FNs and FPs ($p < 0.05$). In this case the dwell on the FNs lasted 1310ms, whereas the dwell on the locations of the FPs was 1970ms.

3.5. *Effect of zoom on performance*

In order to assess if the use of the digital zoom window helped or hurt performance, we have used the first impression provided by the observers, as well as the locations of the clusters with a long visual dwell (>= 1000ms) to score the observers' performance before they were allowed to zoom in on the regions of interest. Thus, for example, if on a lesion-free image the observer had 3 clusters of significant visual dwell, but called the case 'normal' at the end of phase 1, then we scored the observer as having made 3 TNs. On the other hand, if the observer called the same case 'abnormal' at the end of his or her run, then we scored the observer as having made 3 FPs on that image. The same reasoning follows on cases that contained a lesion. Because there was no information available, on phase 1, about the confidence of the observers on their decisions, then ROC analysis could not be used, and we have scored the observers' performance using log odds. Table 4 lists the values for before and after zooming was allowed.

	Initial Impression	After Zooming
NR	Lo = -0.40	Lo = -0.33
R	Lo = -0.36	Lo = 0.88

Table 4. Log odds for the observers performance before and after zooming was allowed.

It can be shown that the gain in the True Positives was about 64% for the R cases, but for the NR cases there was an actual loss of 19% in performance, meaning that the False Positives overtook the True Positives once the use of the zoom window was allowed.

4. Discussion

In this paper we have shown that a digital zoom window can be used to monitor the regions in the image that attracted the observers' attention, as opposed to an invasive infrared eye-tracker. The zoom window was used in the locations of the majority of the decisions made by the observers, even the False Positives and the False Negatives. Furthermore, most zooms occurred in locations where the observers had had long (> 1000 ms) visual dwell. Moreover, the order in which the zooms occurred, as well as the length of the zooms, yielded statistically significant information about decision outcome, which makes the use of a digital zoom window an interesting alternative to aid the observers in improving performance when reading a mammogram test set. Zooming significantly improved performance on the R cases; unfortunately it had the apparent effect of raising the noise level in the cases where a subtle cancerous lesion was present, which decreased performance. Because the test set chosen for this experiment was so heavily biased towards subtle lesions, this decrease in performance was significant. It is unclear if in clinical conditions the use of the zoom window could actually help radiologists to make fewer False Positives and, most importantly, fewer False Negatives.

5. Acknowledgements

This work was partially supported by Grant DAMD17-97-1-7103 between the USAMRMC and C. F. Nodine.

6. References

1. C. F. Nodine, H. L. Kundel, S. C. Lauver and L. C. Toto, "Nature of Expertise in Searching Mammograms for Breast Masses", Academic Radiology, 3:1000-1006, 1996.
2. E. A. Krupinski, C. F. Nodine and H. L. Kundel, "Enhancing Recognition of Lesions in Radiographic Images Using Perceptual Feedback", Optical Engineering, 37:813-818, 1998.